An Integrative Approach to Linguistic Complexity Analysis for German

Dissertation

zur Erlangung des akademischen Grades

Doktor der Philosophie

in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Zarah Leonie Weiß
aus Darmstadt

2024

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekanin:	Prof. Dr. Angelika Zirker
Hauptberichterstatter:	Prof. Dr. W. Detmar Meurers
Mitberichterstatterin:	Prof. Dr. Anke Lüdeling
Mitberichterstatter:	Prof. Dr. R. Harald Baayen
Tag der mündlichen Prüfung:	01.03.2024

Universitätsbibliothek, TOBIAS-lib

für Heidi und Carsten

Acknowledgments

I spend overall ten years at the Department of General Linguistics in Tübingen—first as a bachelor and master student, then as a PhD student and lecturer. In this time, I was supported by my supervisors, collaborators, colleagues, friends, and family. This section is for them.

- First, I would like to thank Detmar Meurers for being my supervisor and mentor throughout the years. Your supervision and advise shaped my work as a researcher and without your support, the late night meetings, and the long discussions, this work would not have been possible.

- I am also grateful to my second supervisor, Anke Lüdeling. Thank you for agreeing to supervise my thesis despite having little time to spare. I am also grateful for your support over the years, first in the LangBank project and later through invitations to your colloquiums and workshops.

- I am grateful to the Lead graduate school and research network which allowed me to engage in interdisciplinary exchange and obtain feedback on my work from a broad audience. I am grateful for the excellent workshops and the platform to network and present my work.

- My profound thanks go to my proof readers: Anita Girelli, Heidemarie Weiß, Heiko Holz, Leona Colling, Marina Haid, Martin Wolf, Ramon Ziai, and Tabea Sanwald. Thank you for your patience and concentrated review. Thank you for keeping me on track, carefully reading the chapters I sent you, and providing so much invaluable feedback. You all did a marvelous job!

- I would also like to thank my colleagues and collaborators from other universities who I had the fortune to get to know over the last years: Anastasia Knaus, Anja Riemenschneider, Barbara Geist, Carolin Odebrecht, Christiane Bertram, Gohar Schnelle, Inga ten Hagen, Jennifer-Carmen Frey, Katrin Wisniewski, Kim Lange-Schubert Lisa Zachrich,

Mareike Kholin, Moritz Sahlender, Pauline Schröter, Stefanie Bredthauer, Stefanie Helbert, Thomas Krause, Theresa Geppert, and Uwe Springmann. Our interdisciplinary exchange has consistently broadened my perspectives and has always brought me great joy. Special thanks go to Hannes Schröter with whom I had the fortune to work with in both, the COLD project and the KANSAS project. Your scientific contributions, organizational talent, and warm support were a great help throughout the years and it was a pleasure working with you.

- I also want to express my gratitude towards my former colleagues in Tübingen: Anja Ehinger, Benedikt Beuttler, Björn Rudzewitz, Çagri Çöltekin, Christl Glauder, Claudia Schulz, David Alfter, Elizabeth Bear, Heiko Holz, Jochen Saile, Johannes Dellert, Kordula De Kuthy, Leona Colling, Maria Morgenstern, Marti Quixal, Ramon Ziai, Sabrina Dittrich, Simón Ruiz, Sowmya Vajjala, Stephen Bodnar, Tanja Heck, and Xiaobin Chen. Thank you for the lively discussions during the numerous coffee and lunch breaks, your feedback in the 'Oberseminar', your company on conferences and work trips, and for always having an open door for me when I had questions or problems no matter how busy you were.

- I would like to thank my spouse, Anita Girelli, for all the support over the years and especially in the last few months. You calmed me down when I felt overwhelmed and listened to me when I rambled excitedly about my research. Your advice on how to be more concise in my writing has been invaluable. Without you, this thesis would have been even longer and, more importantly, my life would be infinitely poorer.

- Special thanks to my friends: Alessandro Greco, Annika Timmins, Heiko Holz, Lars Horber, Leona Colling, Marina Haid, Martin Schiebel, Martin Wolf, Mei Shin Wu, Stefan Brodbeck, Tabea Sanwald, and Tamara Seebacher. Thank you for all the long conversations, game nights, cooking evenings, excursions, parties, laughter, and support. You were by my side in good times and bad times and reminded me that there is more to life than doing research.

- Vorrei anche ringraziare la mia famiglia italiana per l'ospitalità, il sostegno e l'affetto: Irene Girelli, Luigi Girelli, Maria Grazia Tosi, Nicola Girelli e soprattutto Carla Girelli, ed inoltre tutti gli zii e i cugini che sono troppo numerosi per essere nominati qui individualmente, ma che sono inclusi calorosamente.

- I also would like to thank my older brother Maximilian Weiß and his wife (and my friend) Tatiana Weiß. Thank you for your incessant support, love, encouragement, and advise. I am grateful that you were there throughout the years to cheer me up and provide me with a new perspective. I always knew you have my back and I could not ask for more caring and loving siblings.

- Last (but certainly not least), I want to thank my parents, Heidemarie Weiß and Carsten Thielmann. Ohne eure bedingungslose Unterstützung und Liebe wäre ich nicht der Mensch, der ich heute bin. Ihr habt dafür gekämpft, dass ich das Gymnasium besuchen konnte, als niemand sonst daran glaubte, dass ich dort erfolgreich sein könnte. Ihr habt euch ohne zu zögern für meine Bildung aufgeopfert, als ich im öffentlichen Schulsystem nicht weiterkam, und mich auf eine Privatschule geschickt, wo ich die Unterstützung und das Lernumfeld erhalten konnte, die ich brauchte. Als ich mich entschloss, das Medizinstudium abzubrechen, um mich zwischen Theaterwissenschaft und Germanistik zu entscheiden, und in drei Semestern nicht weniger als dreimal das Studienfach wechselte, habt ihr mich immer wieder bestärkt und mir versichert, dass alles gut gehen würde, solange ich meiner Leidenschaft folgte. Ihr hattet natürlich Recht. Ich weiß nicht, wo ich heute wäre, wenn ihr nicht so unermüdlich und bedingungslos an mich geglaubt hättet, aber mit Sicherheit hätte ich weder meine Leidenschaft für Computerlinguistik gefunden noch einen Doktortitel erworben. Diese Arbeit ist euch gewidmet, denn ihr habt den Grundstein gelegt, ohne den all dies nicht möglich gewesen wäre.

Abstract

This thesis develops an integrative approach to automatic linguistic complexity analyses for German and applies it to predict the proficiency of learner writing and the readability of texts for native and non-native speakers of German. Complexity is a central concept in applied linguistics and has been used in Second Language Acquisition (SLA) research to characterize and benchmark language proficiency and to track developmental trajectories of learners (Ortega, 2012). However, the focus of SLA complexity research has been on the analysis of syntax and lexicon and the English language (Housen *et al.*, 2019; Wolfe-Quintero *et al.*, 1998). More research on other linguistic domains—such as morphology or discourse—is needed to model complexity as a multidimensional construct. Furthermore, more languages should be studied to promote complexity research. Measures of linguistic complexity have also been found to be important features in computational linguistic research on Automatic Proficiency Assessment (APA) and Automatic Readability Assessment (ARA). This thesis combines insights from SLA complexity research and computational linguistic approaches to APA and ARA to address important research gaps in SLA complexity research and work on APA and ARA for education contexts.

We propose a linguistically broad approach to complexity that combines measures of syntactic, lexical, and morphological complexity, as well as measures of discourse, human processing, and language use. In doing so, we integrate theories and concepts form different research disciplines including SLA complexity research, computational linguistics, and psychology. We implemented a system to automatically calculate these measures relying on Natural Language Processing (NLP) techniques. With 543 measures, it calculates to the best of our knowledge the largest and most diverse collection of measures of absolute and relative complexity for German. To make this resource accessible to other researchers and thereby promote the comparability and reproducibility of complexity research for German, we integrated this system into the Common Text Analysis Platform (CTAP) by Chen and Meurers (2016). We generalized the originally monolingual web platform for English to support multi-

lingual analyses, leading to its extension to several additional languages. In an empirical study on the impact of non-standard language on the NLP annotations and subsequent calculation of measures, we confirmed that even on language from beginning learners, our analysis remains overall robust and errors hardly impact our complexity estimates or models trained with them.

We then demonstrate the value of our integrative broad linguistic modeling approach to linguistic complexity for APA and ARA. First, we provide an overview of the current research landscape for both domains by conducting two systematic surveys focusing on automatic approaches for German published in the past twenty years. Both surveys showcase the need for more research on approaches targeting second or foreign language (L2) learners and young native speakers, more cross-corpus testing, and more accessible models. For ARA, we observed that traditional readability formulas remain the de facto standard in research that is not specifically dedicated to the development of new ARA approaches, even though they have been criticized as overly simplistic by ARA researchers and generally perform below the current state-of-the-art (SOTA). Second, we report on several machine learning experiments that build on these insights and take into consideration the research needs we identified. We train models for predicting language proficiency for L2 learners on long texts at the full Common European Framework of Reference for Languages (CEFR) scale (A1 to C1/C2) and short answers to reading comprehension questions in the form of course levels (ranging from A1.1 to A2.2). We also train a model for capturing early native language (L1) academic language proficiency of students using grade levels (1st to 8th grade). For text readability, we train models for L2 learners for longer texts (distinguishing texts for learners at the CEFR levels A2, B1/B2, C1) and sentences (using a 7-point Likert scale) as well as a model for German media language aimed at children or adults (making a binary distinction). We test these models across corpora and on hold-out data sets. With this, we illustrate the generalizability of our models across different task contexts, elicitation contexts, languages, and publishers. We also perform linguistic analyses on all data sets studied, which yields important insights into the characterization of developmental trajectories in German. This thesis makes a special methodological contribution to ARA, as we compile a total of three new readability corpora which for the first time facilitate cross-corpus testing and cross-language testing for German ARA.

In sum, this thesis provides novel insights into the developmental variation of linguistic complexity in German and its role for text readability. It also contributes important new resources for research on complexity, ARA, and APA by making available the multilingual CTAP system, new readability corpora, and new models for German.

Kurzfassung

Diese Dissertation entwickelt einen integrativen Ansatz zur automatischen Analyse linguistischer Komplexität für das Deutsche und wendet ihn an, um die Schreibkompetenz von Lernenden und die Lesbarkeit von Texten für deutsche Muttersprachler:innen und Nicht-Muttersprachler:innen vorherzusagen. Komplexität ist ein zentrales Konzept in der angewandten Linguistik und wurde in der Forschung zum Zweitspracherwerb (SLA) verwendet, um die Sprachkompetenz von Lernenden zu charakterisieren und zu messen (Ortega, 2012). Der Schwerpunkt der SLA-Komplexitätsforschung lag hierbei auf der Analyse von Syntax und Lexikon im Englischen (Housen *et al.*, 2019; Wolfe-Quintero *et al.*, 1998). Um Komplexität als multidimensionales Konstrukt zu modellieren, sind weitere Forschungen zu anderen sprachlichen Bereichen erforderlich (beispielsweise Morphologie oder Diskurs). Zudem müssen mehr unterschiedliche Sprachen untersucht werden, um die Komplexitätsforschung voranzubringen. Maße für sprachliche Komplexität haben sich auch in der computerlinguistischen Forschung zur automatischen Sprachkompetenzbewertung (APA) und zur automatischen Lesbarkeitserfassung (ARA) als wichtige Merkmale erwiesen. In dieser Arbeit werden Erkenntnisse aus der SLA-Komplexitätsforschung und computergestützte linguistische Ansätze für APA und ARA kombiniert, um wichtige Forschungslücken in den jeweiligen Disziplinen zu schließen.

Wir schlagen einen linguistisch breit angelegten Ansatz für Komplexität vor, der Maße für syntaktische, lexikalische und morphologische Komplexität sowie Maße für Diskurs, menschliche Sprachverarbeitung und Sprachgebrauch kombiniert. Dabei integrieren wir Theorien und Konzepte aus verschiedenen Forschungsdisziplinen wie der SLA-Komplexitätsforschung, der Computerlinguistik und der Psychologie. Wir haben ein System zur automatischen Berechnung dieser Maße implementiert, das auf Techniken der natürlichen Sprachverarbeitung (NLP) beruht. Mit 543 Maßen berechnet es nach unserem derzeitigen Kenntnisstand die größte und vielfältigste Sammlung von Maßen der absoluten und relativen Komplexität für das Deutsche. Um diese Ressource anderen Forschern zugänglich zu machen und damit die Vergleichbarkeit und Reproduzierbarkeit der Komplexitätsforschung für das Deutsche zu fördern, haben

wir dieses System in CTAP (Chen und Meurers, 2016) integriert. Wir haben die ursprünglich nur für Englisch entwickelte Webplattform generalisiert, um mehrsprachige Analysen zu unterstützen. Dies führte bereits zu ihrer Erweiterung auf mehrere andere Sprachen. In einer empirischen Studie zu den Auswirkungen von nicht-standardisierter Sprache auf die NLP-Annotationen und die anschließende Berechnung der Maße haben wir bestätigen können, dass unsere Analyse selbst bei Sprache von Deutsch-Anfängern insgesamt robust bleibt und etwaige Fehler nur geringe Auswirkungen auf unsere Komplexitätsmessungen oder die damit trainierten Modelle haben.

Im Weiteren demonstrieren wir den Wert unseres integrativen, breit angelegten linguistischen Modellierungsansatzes für linguistische Komplexität für APA und ARA. Zunächst geben wir einen Überblick über die aktuelle Forschungslandschaft für beide Bereiche, indem wir zwei systematische Literaturrecherchen zu automatischen Ansätzen für das Deutsche in den vergangenen zwanzig Jahren durchführen. Beide Erhebungen zeigen den Bedarf an mehr Forschung zu Ansätzen, die sich an Zweit- oder Fremdsprachenlerner und junge Muttersprachler richten, an mehr korpusübergreifenden Tests und an besser zugänglichen Modellen. In Bezug auf ARA stellen wir fest, dass traditionelle Lesbarkeitsformeln weiterhin den Standard in der Forschung darstellen, die sich nicht speziell mit der Entwicklung neuer ARA-Ansätze befasst. Dies ist der Fall, obwohl diese Formeln von ARA-Forschern als zu vereinfachend kritisiert wurden und im Allgemeinen schlechtere Ergebnisse als zeitgenössische Verfahren liefern. Zweitens berichten wir über mehrere Experimente zum maschinellen Lernen, die die von uns so ermittelten Forschungslücken adressieren. Wir trainieren Modelle zur Vorhersage der Sprachkompetenz von L2-Lernern für lange Texte auf der gesamten Skala des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER; A1 bis C1/C2) und kurze Antworten auf Fragen zum Leseverständnis in Form von Kursstufen (von A1.1 bis A2.2). Außerdem trainieren wir ein Modell zur Erfassung der frühen muttersprachlichen akademischen Sprachkenntnisse von Schülern anhand von Klassenstufen (1. bis 8. Klasse). Für die Lesbarkeit von Texten trainieren wir Modelle für L2-Lerner für längere Texte (mit Unterscheidung von Texten für Lerner auf den GER-Niveaustufen A2, B1/B2, C1) und Sätze (unter Verwendung einer 7-Punkte-Likert-Skala) sowie ein Modell für deutsche Mediensprache, das sich an Kinder oder Erwachsene richtet (mit einer binären Unterscheidung). Wir testen diese Modelle über Korpora hinweg und an *Hold-out*-Datensätzen. Damit illustrieren wir die Generalisierbarkeit unserer Modelle über verschiedene Aufgabenkontexte, Erhebungskontexte, Sprachen und Verlage hinweg. Darüber hinaus führen wir für alle untersuchten Datensätze

linguistische Analysen durch, die wichtige Erkenntnisse über die Charakterisierung von Entwicklungsverläufen im Deutschen liefern. Wir leisten dabei einen besonderen methodischen Beitrag zu ARA, indem wir drei neue Lesbarkeitskorpora erstellen, die erstmals die korpus- und sprachenübergreifende Evaluation von ARA-Modellen für das Deutsche ermöglichen.

Insgesamt liefert die vorliegende Arbeit neue Einsichten in die entwicklungsbedingte Variation sprachlicher Komplexität im Deutschen und ihre Rolle für die Lesbarkeit von Texten. Durch die Bereitstellung des mehrsprachigen CTAP-Systems, neuer Lesbarkeitskorpora und neuer Modelle für das Deutsche stellt sie außerdem wichtige neue Ressourcen für die Forschung zu Komplexität, APA und ARA bereit.

List of publications

Parts of this thesis also appear in the following peer-reviewed publications. Chapter 8 includes the full papers.

1. **Weiss, Z.**, Chen, X., & Meurers, D. (2021). Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, 38–54.

2. **Weiss, Z.**, & Meurers, D. (2018). Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, 303–317.

3. **Weiss, Z.**, & Meurers, D. (2019). Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 380–393.

4. **Weiss, Z.,** & Meurers, D. (2019). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research, Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-la-Neuve: Presses Universitaires de Louvain, 419–435.

5. **Weiss, Z.**, & Meurers, D. (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. In *International Journal of Learner Corpus Research*, 7(1), 83–130.

6. **Weiss, Z.**, & Meurers, D. (2022). Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, 141–153.

Funding

The research presented in this dissertation has been funded through the following projects:

- The project “LangBank: Digital Infrastructure to Support the Study of Latin and Historical German” was funded by by the German Research Foundation (DFG) from September 2015 until September 2018, grant number LU 856/7-1, ME 1447/3-1.
- The project “KANSAS: Development of a competence-adaptive, user-oriented search engine for authentic language learning texts” was funded by the Federal Ministry of Education and Research (BMBF) as part of the *AlphaDekade* (engl. “Decade of Literacy”), grant number W143500.
- The project “COLD: Competencies of school teachers and adult educators in teaching German as a second language in linguistically diverse classrooms” was funded by the Leibniz Association as part of the Leibniz Collaborative Excellence funding program from April 2019 to September 2022, grant number K113/2018.

Contents

List of Tables

List of Figures

Acronyms

10-CV 10-folds cross-validation. xxx, 68, 124, 141, 142, 155, 159, 162, 171, 173, 175, 176

AE UIMA Analysis Engine. xxx, 96, 97, 100, 101, 102

AES Automatic Essay Scoring. xxx, 41, 46, 47, 57

ALPS Automatic Language Performance Scoring. xxiv, xxvi, xxx, 5, 8, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 131, 158, 182, 183, 184

AoA age of acquisition. xxx, 19, 20, 30

APA Automatic Proficiency Assessment. xi, xii, xiii, xiv, xv, xxx, 2, 5, 6, 8, 103, 107, 108, 109, 111, 147, 151, 155, 179, 182, 183, 184, 187, 188, 191, 192, 193

API application programming interface. xxx, 193

ARA Automatic Readability Assessment. xi, xii, xiii, xiv, xv, xxiv, xxvi, xxx, 2, 4, 5, 6, 8, 15, 17, 19, 21, 23, 24, 48, 58, 59, 60, 61, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 89, 103, 104, 105, 106, 123, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 147, 149, 150, 151, 165, 166, 167, 169, 170, 172, 173, 174, 175, 176, 179, 184, 185, 186, 187, 188, 191, 192, 193

ATS Automatic Text Scoring. xxx, 2, 8, 13, 15, 41, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 69, 73, 83, 89, 103, 107, 108, 109, 111, 118, 119, 121, 127, 137, 182, 189, 193

AWE Automatic Writing Evaluation. xxx, 49, 50, 55, 56, 182

BICS Basic Interpersonal Communication Skills. xxx, 37, 44, 45, 61, 66

BLC Basic Language Cognition. xxx, 45

CA conversational agents. xxx, 71

CAF Complexity, Accuracy, and Fluency. xxx, 10, 11, 28, 35, 37, 38, 39, 42, 55, 57, 82, 119, 148, 151, 158, 161

CALP Cognitive Academic Language Proficiency. xxx, 37, 44, 45, 61, 66

CAS UIMA Common Analysis Structure. xxx, 95, 96, 101, 102

CEFR Common European Framework of Reference for Languages. xii, xxx, 3, 43, 53, 67, 68, 73, 78, 79, 115, 116, 148, 152, 155, 157, 168, 182, 186

CoNLL Conference on Computational Natural Language Learning. xxx, 91

CREG Corpus of Reading comprehension Exercises in German. xxx, 153, 162, 165

CSV comma-separated value. xxx, 90, 102

CTAP Common Text Analysis Platform. xi, xii, xiv, xv, xxiv, xxvi, xxx, 4, 87, 88, 89, 94, 95, 96, 97, 98, 100, 101, 102, 127, 128, 142, 145, 173, 175, 181, 182, 193, 387, 388, 389, 390, 392, 394, 396, 397, 399, 402, 407, 408, 410, 411, 412, 413, 414, 415, 417, 418, 422

DLT Dependency Locality Theory. xxx, 27, 94

ESL English as a Second Language. xxx, 29

FLA First Language Acquisition. xxx, 21, 24, 31, 34

GWT Google Web Toolkit. xxx, 95

HD-D Hypergeometric Distribution Diversity. xxx, 18, 397

HLC Higher Language Cognition. xxx, 45

IRR inter-rater reliability. xxx, 49, 52, 53, 121, 137

KCT Karlsruhe Childrens' Text. xxiv, xxx, 93, 154, 158, 160

KU University of Kansas. xxx, 153, 165

L1 native language. xii, xxx, 1, 2, 3, 5, 6, 7, 8, 10, 12, 14, 19, 20, 22, 23, 24, 29, 31, 32, 33, 35, 36, 40, 42, 44, 45, 58, 61, 62, 69, 70, 73, 78, 112, 113, 116, 120, 126, 134, 147, 148, 149, 152, 157, 158, 159, 165, 169, 172, 174, 182, 183, 184, 185, 188, 189, 191, 192

L2 second or foreign language. xii, xiv, xxiv, xxx, 1, 2, 3, 5, 6, 7, 8, 10, 12, 14, 17, 19, 22, 23, 26, 28, 29, 30, 31, 32, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 51, 55, 58, 61, 66, 67, 68, 69, 70, 73, 74, 77, 78, 79, 88, 111, 112, 114, 115, 116, 120, 123, 124, 126, 130, 134, 145, 147, 148, 149, 150, 151, 152, 153, 155, 156, 157, 158, 161, 162, 165, 168, 172, 174, 175, 181, 182, 183, 184, 185, 186, 188, 192

LIX Läsbarthetsindex. xxx, 68, 70

LSA Latent Semantic Analysis. xxx, 20, 25, 57, 72

MAE mean absolute error. xxx, 52, 73

MCI Morphological Complexity Index. xxv, xxx, 17, 22, 411

MOOC Massive Open Online Course. xxx, 49

MTLD Measure of Textual Lexical Diversity. xxx, 18, 94, 397

NLG Natural Language Generation. xxx, 71, 83

NLP Natural Language Processing. xi, xii, xiii, xiv, xxiv, xxx, 2, 3, 51, 71, 72, 79, 80, 87, 88, 90, 91, 92, 93, 95, 97, 98, 99, 100, 101, 102, 148, 154, 162, 164, 165, 181, 183

NP noun phrase. xxx, 16, 30, 93, 94, 384, 390, 391, 393

NT non-terminal. xxx

ORF Ordinal Random Forest. xxx, 162, 173

OSU The Ohio State University. xxx, 153

PCFG probabilistic context-free grammar. xxx

PID propositional idea density. xxx, 25, 94

POS part-of-speech. xxx, 18, 22, 25, 31, 80, 89, 90, 92, 93, 98, 99, 101, 162, 164, 394, 397, 398, 403, 404, 406

PP prepositional phrase. xxx, 384, 385, 390, 394

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses. xxx, 5, 104

PTB Penn Treebank. xxx, 91

RMSD root mean squared difference. xxx, 163, 164

RMSE root mean squared error. xxx, 52, 73, 74, 79, 82, 121

SLA Second Language Acquisition. xi, xiii, xxx, 1, 2, 7, 8, 9, 10, 11, 13, 14, 15, 18, 21, 23, 24, 26, 28, 31, 34, 38, 40, 41, 42, 50, 55, 56, 82, 87, 88, 89, 106, 111, 112, 149, 151, 155, 179, 180, 181, 182, 183, 187, 193

SMO Sequential Minimal Optimization. xxx, 141, 155, 159, 170, 171

SOTA state-of-the-art. xii, xxx, 5, 56, 67, 68, 72, 74, 79, 82, 83, 85, 104, 123, 124, 127, 128, 140, 141, 143, 144, 145, 176, 179, 184, 193

TF-IDF term frequency-inverse document frequency. xxx, 81

TTR type token ratio. xxx, 17, 18, 92, 123, 397, 398

UIMA Unstructured Information Management Application. xxx, 95, 96, 100, 101, 102

UTF-8 Universal Transformation Format-8. xxx, 92

VP verb phrase. xxx, 385, 390, 391, 393

XML Extensible Markup Language. xxx, 96, 100, 101

ZPD Zone of Proximal Development. xxx, 60, 67, 192

Chapter 1

Introduction

1.1 Aims and contributions

This dissertation develops a comprehensive and integrative approach to automatic linguistic complexity modeling for German and illustrates the highlights of this approach by predicting the proficiency of L1 and L2 learners based on free writing samples as well as the readability of long and short texts for L1 and L2 speakers of German. Linguistic complexity is a prolific construct in linguistics: It has been used to compare and describe languages and to study language use in theoretical linguistics (e.g., Biberauer *et al.*, 2014; Chomsky, 1956; Trotzke and Zwart, 2014) as well as dialectology, typology, and historical linguistics (e.g., Miestamo *et al.*, 2008, and articles therein)—here with a focus on diachronic and synchronic variation. Complexity is also a relevant research subject in psycho-linguistics (e.g., Dussias, 2001; Menn and Duffield, 2014; Shain *et al.*, 2016; Wendt *et al.*, 2014) and computational linguistics (e.g., Brunato *et al.*, 2016, and articles therein). One of the most advanced theoretical approaches to complexity research comes from SLA research. In SLA research, linguistic complexity has been used to estimate (L2) proficiency, characterize language performance, and to benchmark the development of learners (Ortega, 2012, p. 128). Coming from research on task-based language learning and teaching, also task variation has been recognized as a relevant factor. Task variation plays an important role in addition to and interacting with developmental variation in SLA complexity research (e.g., Alexopoulou *et al.*, 2017; Pallotti, 2019; Staples *et al.*, 2016; Tavakoli and Foster, 2011). However, linguistic complexity in SLA research has so far primarily been studied in the domains of syntax and lexicon. This has been criticized as reductionist (Housen *et al.*, 2019), because it ignores other linguistic domains that are important for language development such as morphology which has only recently entered

the focus of SLA complexity research. This holds especially for languages other than English, which are under-represented in SLA complexity research.

The measures studied in SLA complexity research have independently (but in parallel) proven to be highly informative in computational linguistic research on Automatic Text Scoring (ATS) (Crossley, 2020; Uto, 2021) and Automatic Readability Assessment (ARA) (Collins-Thompson, 2014; Vajjala, 2022). These feature-based machine learning approaches continue to be relevant despite the increasing relevance of deep learning approaches: Feature-based machine learning using theoretically motivated features supports interpretable predictions. Interpretability is especially important in educational contexts (Attali and Burstein, 2006, p. 6; Powers *et al.*, 2002, p. 2; Zhang, 2013, p. 13). Consideration of linguistically broad evidence has been found to be particularly robust and informative in machine learning approaches (e.g., Vajjala, 2018). The computation of a large number of features is supported by the use of NLP tools, which enable the fully automatic identification of linguistic constructions. Automating the calculation of features not only makes the consideration of a broad range of constructs feasible. It also has the potential to foster the comparability and reproducibility of findings (Crossley and McNamara, 2014; Lu, 2010; Ströbel *et al.*, 2020, p. 738). Accordingly, many systems have been developed for the study of the English language (e.g., Chen and Meurers, 2016; Crossley *et al.*, 2016c; McNamara *et al.*, 2010a). However, comparatively little research has been dedicated to the automatic analysis of linguistic complexity in German or generally languages other than English.

The aim of this dissertation is to address these research gaps in SLA complexity research and computational linguistic work on APA and ARA. To do so, I present a linguistically uniquely broad approach to automatic complexity analysis that can be flexibly extended for other languages. Furthermore, I demonstrate the applicability of this linguistically broad approach to APA and ARA, filling important research gaps for German in these computational linguistic research areas. I empirically identify these research gaps through two systematic literature surveys which are respectively the first systematic surveys for APA and ARA for German. The core contributions of this dissertation fall into three categories:

1. The generation of research resources including: three readability corpora for German, two systems for the automatic broad linguistic complexity analysis of German, and several models for the prediction of readability and proficiency for L1 and L2 learners.

2. Methodological contribution to SLA complexity research and computational linguistic research on APA and ARA including promoting the assessment of cross-corpus, cross-

task, and cross-language generalizability of complexity modeling approaches as well as the assessment of how robust NLP and automatic complexity assessment are on non-standard data.

3. Linguistic insights into the developmental variation of complexity in L2 German writing across the full CEFR range and the first years of L1 academic language development as well as into the adaptation of reading materials for L1 and L2 readers across different publishers.

1.2 Overview

The remainder of this thesis is divided into five content chapters (Chapter 2 to Chapter 6), followed by the bibliographies (Chapter 7), publications (Chapter 8) and the appendices (Chapters A and B). In the following, I briefly outline the contents of the five content chapters.

Background (Chapter 2) The background chapter introduces the central concepts, methods, and research domains that are relevant for the work presented in this thesis. Section 2.1 focuses on the construct 'linguistic complexity' and on applied linguistic research on linguistic complexity. In Section 2.1.1, I provide a definition of complexity that distinguishes between absolute and relative complexity, an important distinction throughout this thesis. I then elaborate on the different dimensions of absolute and relative complexity that have been studied in applied linguistic research (Section 2.1.2). There, I highlight the necessity for taking a broad linguistic perspective on complexity to account for interactions and variational differences between complexity (sub-)domains. Finally, I discuss the role of variation for linguistic complexity research (Section 2.1.3). I focus on developmental variation, task and register variation, and cross-lingual variation, because these are central for the empirical studies that are part of this thesis. I argue for taking into account the variation in complexity caused by different sources (especially developmental and task variation), rather than considering them in isolation. This is reflected in the empirical studies in this dissertation, in which I systematically focus on task generalization of models of language proficiency.

Section 2.2 is dedicated to computational linguistic work on the automatic assessment of language proficiency. First, I define 'language proficiency' (Section 2.2.1), focusing on perspectives from language testing research. Second, I discuss common application domains of automatic proficiency assessment (Section 2.2.2) to highlight the relevance of this task within

and beyond education. Third, I elaborate on the current methods and trends in computational linguistic research on automatic proficiency assessment (Section 2.2.3). This section focuses on supervised machine learning approaches to automatic proficiency assessment and discusses central concepts and ongoing methodological challenges. The considerations presented in this section play an important role in the design of the systematic literature survey on proficiency assessment (Section 4.2) as well as the empirical studies on automatic proficiency assessment from this dissertation (Section 5.2).

Section 2.3 provides a comprehensive background on computational linguistic research on ARA. Its structure is similar to Section 2.2. It starts with a definition of readability (Section 2.3.1) that considers computational linguistic and psychological or psycho-linguistic research on discourse comprehension. I then discuss different application domains for ARA (Section 2.3.2) to illustrate its interdisciplinary relevance. Afterwards, I discuss the current methods and trends in computational linguistic research on automatic readability assessment (Section 2.3.3), again with a focus on supervised machine learning. The considerations presented in this section play an important role in the design of the systematic literature survey (Section 4.3) as well as the empirical studies on ARA from this dissertation (Section 5.3). To avoid redundancies between Section 2.2.3 and Section 2.3.3, Section 2.3.3 partially references back to concepts and methodological considerations that have already been discussed in detail in Section 2.2.3. Section 2.3 concludes with a brief remark on readability formulas. Here I focus on contrasting the role of readability formulas in research on and applications of ARA.

Automating German complexity modeling (Chapter 3) This chapter describes the technical details of the two complexity analysis systems that I developed and used in this dissertation. Section 3.1 starts with a general overview that motivates the relevance of automatic complexity modeling for applied linguistics in general as well as the need for a (multi-lingual) system for the analysis of the German language. This section also outlines the conceptual analysis workflow that both systems share. I then describe the technical details of the monolingual German legacy system (Section 3.2) which I used throughout this thesis in all but two studies (i.e., Weiss and Meurers, 2022; Weiss *et al.*, 2021). Section 3.3 provides the technical details of the mutlilingual CTAP system which is a web-based analysis platform that currently supports the analysis of German, English, French, Dutch, Spanish, and Portuguese. The section focuses on the German component and the general architecture of the multilingual analysis pipeline which I introduced to the originally monolingual English platform in the context of

this thesis. Other languages are discussed where it serves the illustration of the multilingual design. This chapter presents the latest developmental status of the two systems at the time of writing. Since the development of both systems was part of the dissertation project, most studies use older versions of the systems. The specific resources used for individual studies can be found in the system descriptions of the respective papers.

Systematic literature surveys (Chapter 4) To complement the general research overview on APA and ARA, I conducted two systematic surveys that characterize research in both application domains for German in the past twenty years. Both surveys were based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement to ensure transparency and completeness. Their shared design principles are introduced in Section 4.1. The specific research questions that I aimed to address with the surveys focus on i) the spread of APA and ARA approaches and methods across research disciplines, ii) for who APA and ARA is being used, and iii) which machine learning methods and which linguistic features are used. I further characterize the current SOTA performance for German and evaluate the availability and accessibility of SOTA methods.

Section 4.2 presents the survey for APA.[1] It shows that despite a steady increase in publications over the observed time period, there is still little research matching the previously defined study criteria. The survey further demonstrates that machine learning-based approaches dominate the field but that more cross-corpus validation is needed to ensure the generalizability of approaches. Most research has focused on the holistic assessment of adults and essay writing, demonstrating the need for more work on young writers and different text types.

Section 4.3 presents the survey for ARA. It shows that ARA has been applied across a broad range of research disciplines. It also highlights that outdated readability formulas remain the de facto standard in research with the sole exception of computational linguistic work specifically dedicated to developing new ARA approaches. The survey further demonstrates that more cross-corpus validation is needed to ensure the generalizability of approaches. Most research has focused on readability for adult L1 readers and long text passages, demonstrating the need for more work on L2 and young readers as well as shorter texts.

Foundational complexity research (Chapter 5) Chapter 5 starts with a section discussing and linking the core contributions of all subsequent papers (Section 5.1). Afterwards, the core

[1] Throughout the survey, I use the broader term Automatic Language Performance Scoring. I motivate this terminological shift in more detail in Section 1.3 (p. 8) and Section 4.1 (p. 103).

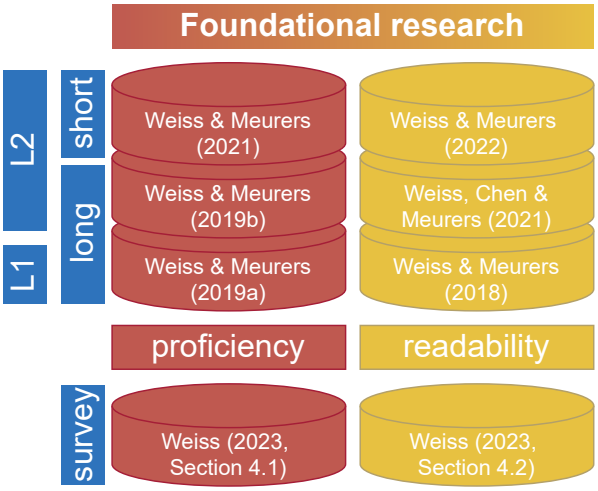


Figure 1.1: *Structured overview of empirical studies presented in this thesis*

methods and findings of each paper are summarized. Section 5.2 summarizes all papers on APA. Section 5.3 summarizes all papers on ARA. In both sections, I first introduce the corpora that I used before elaborating on the individual papers. The papers themselves can be grouped along the two dimensions of target language (L1 versus L2) and sample length (long versus short) as illustrated in Figure 1.1. Note that each of the six papers contributing to this thesis is presented at three levels of granularity: a three-sentence summary in the frame (to obtain an overview that is situated within the broader scope of this thesis), an extended abstract in the subsection dedicated to the paper (as a summary of the methods and main findings), and the full paper in Chapter 8 (for a complete and fully self-contained discussion).

Conclusion (Chapter 6) The conclusion of this thesis is divided into three sections: Section 6.1 summarizes the limitations as well as the major contributions and achievements of this thesis. I focus on the questions of resources, methodological advancements, and linguistic findings. Section 6.2 discusses the implications of this work for interdisciplinary research and teaching practice. I here outline how my integrative approach to broad linguistic modeling can be used to analyze data from authentic education contexts and to foster collaborations with other disciplines such as education science and history didactics. I end with an outlook on future research directions that can build on the presented work in Section 6.3.

1.3 Notational and terminological conventions

I use the following terminological and notational conventions throughout this dissertation.

- I use special notations to denote 'concepts', `names of complexity measures or features`, and *emphasis*.

- I use the terms 'complexity measure' and 'complexity feature' in this thesis. The notion of measures is more prevalent in SLA complexity research. The notion of features is used in machine learning. I use the terms in this thesis correspondingly, using 'measures' unless I refer to a machine learning context.

- Even though I predominantly use the first person singular ('I') throughout this thesis, I also occasionally use 'we', either to include the reader or to highlight that I make reference to collaborative efforts.

- The notions 'corpus' and 'data set' are often used interchangeably when working with language data (e.g., McCarthy and Jarvis, 2013, p. 48). Corpora are systematic collections of language data that represent some form of authentic language use for a specific language variety (Lüdeling and Kytö, 2008, p. iv; McCarthy and Jarvis, 2013, p. 48). The notion 'data set' can also apply to non-language data and is also used outside of linguistics. In this thesis, I use both notions interchangeably, treating 'corpus' as a hyponym of 'data set'.

- I use the distinction between L1 speakers and L2 speakers throughout this thesis. This distinction has a long tradition in linguistic research. However, the notion of L1 speakers should be used with care because it has often been used to imply an ideal type of language use. This is a misconception that is attracting increasing criticism (e.g., Birdsong and Gertken, 2013; Bonfiglio, 2010; Doerr, 2009). First, researchers have warned against tapping into the 'comparative fallacy' (Bley-Vroman, 1983) of measuring L2 performance in terms of its proximity to L1 performance. Second, the distinction ignores the considerable inter-individual heterogeneity of native speakers' language use (Shadrova *et al.*, 2021). Third, turning to school contexts, Pallotti (2017, p. 400) pointed out that the distinction fails to account for the fact that pupils are often L1 speakers of a dialect rather than the standard variety of a target language and that the acquisition of the standard variety parallels L2 acquisition (see also Siegel, 2010). In this thesis, I use

the term L1 as a broad categorization of a target group of advanced language learners for APA and ARA, without subscribing to its use as a homogeneous reference norm for ideal language use. Where appropriate, I deconstruct L1 and L2 proficiency further into competencies for different language codes (following Pallotti, 2017, p. 401).

- I use the term 'language learner' throughout this thesis. Unless specified differently, this jointly refers to L2 learners as well as L1 learners.

- As you read this dissertation, you will notice that I use several terms to refer to the automatic prediction of language proficiency. Mainly I use the term Automatic Proficiency Assessment (APA). This connects to the SLA tradition for studying linguistic proficiency (Ortega, 2012, p. 128) and the term is also used in other computational linguistics work on this topic (e.g., Bannò and Matassoni, 2022; Metallinou and Cheng, 2014; Vajjala and Lõo, 2013). However, I deviate from this terminology in two places for content reasons. First, in the background section on computational linguistics research (Section 2.2), I use the term Automatic Text Scoring (ATS) because APA is a sub-type of this line of research and I discuss work on ATS in general in Section 2.2 for the sake of completeness. Second, also in the systematic survey on automatic assessment of linguistic proficiency (Section 4.2), the term APA proved to be too narrow, as my survey criteria also consider related approaches to automatic evaluation of linguistic performance. At the same time, ATS is too general because it lacks the focus on *language* performance. I propose the term Automatic Language Performance Scoring (ALPS) as a generic term for the survey chapter, which I motivate in more detail in Section 4.1 (p. 103).

Chapter 2

Background

2.1 Linguistic complexity research

Linguistic complexity is a highly productive construct that has been used to characterize and compare languages and language use. It has been utilized across a broad variety of linguistics subdisciplines. For example, in SLA research, linguistic complexity has been used to characterize developmental trajectories and proficiency differences between learners (e.g., Housen *et al.*, 2012, 2019; Kuiken *et al.*, 2019; Norris and Ortega, 2009; Wolfe-Quintero *et al.*, 1998). Formal theoretical linguistics has used the notion of complexity to compare linguistic structures under the perspective of specific (often generative) linguistic theories (e.g., Biberauer *et al.*, 2014; Chomsky, 1956; Trotzke and Zwart, 2014). Also functional and contact-based linguistic research has utilized the notions 'complexity' and 'simplicity' to explain diachronic and synchronic language change in dialectology, typology, and historical linguistics (e.g., Miestamo *et al.*, 2008, and articles therein). Linguistic complexity has also been studied in psycho-linguistics (e.g., Dussias, 2001; Menn and Duffield, 2014; Shain *et al.*, 2016; Wendt *et al.*, 2014) and computational linguistics (see Brunato *et al.*, 2016, and articles therein as well as Sections 2.2 and 2.3).

The emphasis of this section is on the SLA perspective on linguistic complexity. This thesis focuses on complexity analyses for education contexts, which makes SLA complexity research a particularly suited conceptual foundation. Furthermore, SLA complexity research provides a theoretically, methodologically, and empirically rich research background for the present work. However, in the interest of an integrative and linguistically broad approach to complexity, approaches from other (sub-)disciplines are referenced as well throughout this section. The remainder of this section is organized as follows: Section 2.1.1 provides an

overview of the efforts to define and operationalize linguistic complexity as a construct. Section 2.1.2 elaborates on several dimensions of linguistic complexity that played an important role in complexity research. Section 2.1.3.1 introduces prior research on variation in linguistic complexity. It focuses on variation due to language development and proficiency, task and register variation, as well as cross-linguistic variation.

2.1.1 What is linguistic complexity?

2.1.1.1 Complexity, Accuracy, and Fluency

In SLA research, complexity has been used for decades as one of three dimensions that characterize language (or specifically L2) performance, the other two dimension being accuracy and fluency. This is commonly referred to as the Complexity, Accuracy, and Fluency (CAF) triad (Housen and Kuiken, 2009; Housen *et al.*, 2012; Skehan, 1998; Wolfe-Quintero *et al.*, 1998). Before turning to the definition of complexity, let us briefly review the concepts of accuracy and fluency. According to Housen and Kuiken (2009, p. 461), accuracy and fluency have been used to analyze (spoken) L2 productions in instructed settings since the 1980s. Accuracy is commonly defined as conformity to target norms (Pallotti, 2009; Wolfe-Quintero *et al.*, 1998) or error-free language production (Lennon, 1990). It is typically measured in terms of various error rates (Housen and Kuiken, 2009; Pallotti, 2009; Wolfe-Quintero *et al.*, 1998). Studies of accuracy risk falling into the 'comparative fallacy' (Bley-Vroman, 1983) due to their inherent need to view language learners' inter-language system (Selinker, 1972, pp. 214–215) in terms of its (dis)similarity to a native language norm. This makes accuracy measures prone to obscure inherent systematicities in the interlanguage system that SLA researchers seek to describe. For this reason, the validity of accuracy measures to track L2 development has been called into question (e.g., Pallotti, 2009, 2017; Wolfe-Quintero *et al.*, 1998). However, they have shown to be useful predictors of (advanced) L2 and L1 development (Norris and Ortega, 2003, p. 737). As a construct, accuracy has been praised to be well defined and "perhaps the simplest and most internally coherent construct" (Pallotti, 2009, p. 592) in the CAF triad, see also Housen and Kuiken (2009, p. 463).

Fluency refers to the capacity to produce uninterrupted speech at a native-like speed (Ellis, 2003, p. 342; Skehan, 2009, p. 510) and native-like processing speed (Lennon, 1990, p. 390). Commonly measured sub-dimensions of fluency include breakdown fluency (measured through pause-based metrics), repair fluency (measured through correction metrics),

and speed fluency (measured through time-based metrics), for an overview, see Tavakoli and Skehan (2005, pp. 254–255). Fluency research has focused on spoken language (cf. Housen and Kuiken, 2009, p. 463) and is less commonly applied to written language productions. As with accuracy, the definition of fluency assumes a normative ideal degree of fluency which should be reached but not exceeded.

Since its addition to the triad in the 1990s (cf. Housen and Kuiken, 2009), complexity has attracted substantial research interest (for overviews, see Housen *et al.*, 2012, 2019; Wolfe-Quintero *et al.*, 1998). Yet, its definition continues to lack clarity. Pallotti (2009, p. 592) calls complexity "the most problematic construct of the CAF triad". He reasons that this is due to the polysemy of the word 'complexity', see also Hennig (2017), Housen and Kuiken (2009), and Weiss (2017) for a similar argument. Housen *et al.* (2019) found that there is no universally accepted definition of complexity and that the research landscape continues to be characterized by an ever-growing diversity of operationalizations and complexity measures, see also Mitchell (2009) for a similar assessment a decade earlier. This might partially be due to the different underlying research questions and theoretical frameworks used in complexity research across linguistic fields that were mentioned at the beginning of this section. That being said, complexity research uses certain definitions and distinctions that re-occur across linguistic fields. In his philosophical overview on complexity, Rescher (1998) defines complexity as follows (emphasis added by me): "Complexity is first and foremost a matter of the *number* and *variety* of an item's constituent elements and of the *elaborateness of their interrelational structure* [...] Any sort of system or process—anything that is a structured while consisting of interrelated parts—will be to some extent complex" (Rescher, 1998, p. 1). This general quantitative notion of complexity has been widely received in complexity research, for example in SLA (Bulté, 2013; Housen *et al.*, 2019) and linguistic typology (Karlsson *et al.*, 2008). Similarly, Ellis defines linguistic complexity as "[t]he extent to which [...] language [...] is elaborate and varied" (Ellis, 2003, p. 340). In this context, elaborateness refers to the number of elements and variation to the number of different types of elements. This notion of complexity directly focuses on language and is one of the dominating definitions in (SLA) complexity research.

2.1.1.2 Types of complexity

Following these general definitions, a broad range of complexity measures has been proposed to quantify the elaboration and variation of language. The definition was applied both glob-

ally to the entire language system and locally to different linguistic (sub-)domains within the theoretical linguistic system, such as syntax, lexicon, and morphology (e.g., Housen *et al.*, 2012; Miestamo, 2008). The operationalization of global complexity measures is notoriously challenging and has even been characterized as "a probably hopeless endeavor" (Kortmann and Szmrecsanyi, 2012, p. 8). As Miestamo (2004, 2008) points out, individual measures of global complexity struggle to be *representative* for the entire language system because in practice they necessarily miss certain aspects. At the same time, a single global complexity metric would have to *compare* and weigh local aspects of the language system against each other to obtain a meaningful aggregate measure. Some researchers have used information theoretic approaches based on computational compression rates to measure global complexity (e.g., Ehret, 2018; Ehret and Szmrecsanyi, 2016, 2019; Juola, 2008), see also Section 2.1.2.7. Operationalizations for measures of local complexity have been more common across research fields. For example, clausal elaboration has been commonly associated with measures estimating the degree of clausal subordination and coordination (Ortega, 2003; Wolfe-Quintero *et al.*, 1998), see also Section 2.1.2.1. Lexical variation has been measured predominantly in terms of lexical diversity (Jarvis, 2013; Lu, 2012), see also Section 2.1.2.2.

Also word frequency measures have been commonly associated with lexical complexity, more specifically lexical sophistication (Lu, 2012; Wolfe-Quintero *et al.*, 1998). They have proven to be highly informative for the evaluation of language performance (see Section 2.1.2.2). However, they do not operationalize the same notion of complexity as the previous examples. Instead of quantifying the local elaborateness and variation of the theoretical linguistic system, frequency measures are motivated by psycho-linguistic insights into human language use and processing. Rare vocabulary requires higher processing times and is often acquired later by native speakers. To distinguish between these two notions of complexity, the distinction between 'absolute complexity' and 'relative complexity' has been proposed. Using the same terminology as Housen *et al.* (2012) and Miestamo (2004), absolute complexity refers to measures of the inherent properties of a language system or language sample following the definition by Rescher (1998) and Ellis (2003). Relative complexity refers to measures that quantify linguistic constructions that are cognitively demanding or acquired at later developmental stages—typically for specific groups such as young L1 learners or adult L2 learners, but potentially also for individuals. These measures are not tied to inherent properties of the language but to aspects of human language use and processing. The contrast between these two types of complexity has been adopted across linguistic sub-disciplines in-

cluding SLA (e.g., Bulté and Housen, 2014; Housen *et al.*, 2012, 2019) and typology (Kusters, 2008; Miestamo, 2008). However, other or additional terminology is used for this distinction in some cases. Kusters (2008) proposed the notion of 'outsider complexity' to discuss the relative complexity of a linguistic domain (e.g., morphology) for different types of 'universal outsiders'. Pallotti (2015a) suggested to use the notion of 'difficulty' instead of 'relative complexity' to emphasize the difference between these two concepts. A similar argument is made by Dahl (2004, pp. 39–40), who proposed to distinguish complexity (meaning absolute complexity) from subjective concepts such as 'cost', 'difficulty', and 'demandingness' for sake of terminological clarity. These are only a few examples of the different terminologies used. Following Housen *et al.* (2012), I use the terms 'absolute complexity' and 'relative complexity' throughout this thesis in the interest of an integrative notion of linguistic complexity.

These terminological distinctions have direct practical implications for complexity research. Several researchers have pointed out how maintaining a conceptually well-motivated basis of measures in complexity research promotes the validity and interpretability of findings. Pallotti (2009) cautions against exclusively considering measures that maximize the variance explained between target populations because this risks tapping into the 'necessary variation fallacy' (Pallotti, 2009, p. 590). He uses this term to refer to the fact that a lack of variation can be as meaningful as the observation of systematic variation, at least when measuring a conceptually relevant aspect of the construct under investigation (see also Norris and Ortega, 2009). In language testing and assessment research, this phenomenon has been discussed under the notion of 'under-representation' and was identified as detrimental to the construct validity of an assessment (Messick, 1996, p. 244). Closely related to this concept—and another potential risk of the purely data driven selection of measures without sound conceptual underpinnings—is the notion of 'construct irrelevance' (Messick, 1996, p. 244). A measure can be highly successful in distinguishing between different groups of interest—such as language learners at different proficiency levels—while being irrelevant to the underlying construct—here language proficiency. A well established example for construct irrelevance in writing quality assessment is `text length`. I discuss the notions of construct validity, construct under-representation, and construct irrelevance again in more detail in Section 2.2.3.1 (pp. 53–54), where I focus their relevance for computational linguistic research on ATS.

Beyond these concerns relating to the validity of studies, insufficiently motivated measures can also hinder the interpretation of findings (Bulté and Housen, 2014; Norris and Ortega, 2009). This becomes particularly evident in the literature analysis by Bulté and Housen

(2014). They reported that SLA complexity research has often used the notion of complexity interchangeably with later L1 acquisition, advanced L2 development, higher proficiency, lower frequency, higher difficulty or 'better' language use (Bulté and Housen, 2014, p. 46). This loose equation of concepts is problematic and promotes circular reasoning. The link between complexity and language development is one of the central topics of investigation in complexity research (see Section 2.1.3.1). An a priori equation of complexity and development is therefore misleading. It is also inappropriate seeing that there are communicative situations in which L2 learners produce more complex utterances than L1 speakers (Pallotti and Ferrari, 2008). This example illustrates that the equation of 'more complex' and 'better' not necessarily holds across communicative contexts and registers (see also Ferrari, 2012; Kusters, 2008; Ortega, 2003; Pallotti, 2009). Any approach to measure linguistic complexity should therefore carefully consider which measures to include as well as their underlying motivation to promote an informed interpretation findings.

2.1.2 Dimensions of linguistic complexity

Linguistic complexity is a multi-faceted construct that has been conceptualized and operationalized from a broad range of research perspectives. A variety of global and local complexity measures has been proposed ranging from the domains of syntax, lexicon, semantics, morphology, and phonology to measures of discourse and information processing. These complexity domains (e.g., syntax and lexicon) as well as sub-domains (e.g., clausal and phrasal complexity) have been shown to interact and vary independently from each other based on factors such as tasks and proficiency, see for example Norris and Ortega (2009, pp. 562–564) and Section 2.1.3 for details. This makes it important to track complexity through a wide range of measures to obtain stable and generalizable estimates (see also Biber *et al.*, 2016; Bulté and Housen, 2014; Lu, 2011). Despite the abundance of complexity measures that has been proposed, only a selected few of them have been used repeatedly and systematically across studies, as demonstrated by the surveys by Ortega (2003) for syntactic complexity and by Bulté and Housen (2012) for complexity measures across linguistic dimensions. This makes it difficult to understand which results from studies can be generalized and which are idiosyncratic (Lu, 2011, pp. 37–38; Ortega, 2003, pp. 494–495). Also, most complexity measures have focused on the domains of syntax and lexicon. This narrow perspective on complexity has been criticized as reductionist (Bulté and Housen, 2012; Housen *et al.*, 2019). More work on other domains of complexity has started to emerge (e.g., Brezina and Pallotti, 2019; Ehret

and Szmrecsanyi, 2019; Paquot, 2019), but the research gap between syntax and lexicon on the one hand and other complexity domains on the other is far from closed. In the following, I introduce the complexity domains that play a relevant role in complexity research taking into consideration both, measures of absolute complexity and measures of relative complexity.

2.1.2.1 Syntactic complexity

Syntactic complexity has arguably been the most intensively studied domain of absolute complexity (Bulté and Housen, 2012, p. 34; Kuiken *et al.*, 2019, p. 162). Absolute syntactic complexity measures have a long history in SLA complexity research (Ortega, 2003; Wolfe-Quintero *et al.*, 1998). They focus on the syntactic elaboration and variation that a language sample exhibits. Measures of syntactic complexity have also been extensively used in work on ATS (for an overview, see Crossley, 2020) and ARA (for an overview, see Collins-Thompson, 2014). Three types of absolute syntactic complexity measures are commonly employed: a) global syntactic complexity measures, b) clausal complexity measures, and c) phrasal (or sub-clausal) complexity measures (Norris and Ortega, 2009, pp. 558–560).

Predominantly length-oriented measures such as `number of words per sentence` are typically considered global measures of syntactic complexity because they can be influenced by any number of clausal or phrasal processes (Norris and Ortega, 2009, p. 561). Other global syntactic complexity measures are based on constituency tree structures and include metrics such as the `terminal to non-terminal node ratio`. Also, `number of words per t-unit` has been a popular measure of global syntactic complexity (Ortega, 2003). A t-unit is the "minimal terminable unit" (Hunt, 1965, p. 305–306) on a sentential level. It consists of a main clause and all of its dependent clauses and embedded clausal structures (Hunt, 1970). The t-unit thus differs from a sentence in two main points: First, sentences can contain multiple main clauses, but t-units cannot (see also Hunt, 1970, p. 199). Second, sentences are commonly defined in terms of graphematic markers and become notoriously ill-defined in their absence (see discussion in Schmidt, 2016). In contrast, t-units are defined syntactically and do not rely on graphematic markers. This promoted t-units as a standard unit of measurement for syntactic complexity measures of speech and writing (Crossley, 2020; Foster *et al.*, 2000; Lu, 2011; Ortega, 2003), despite ample criticism of the unit and its partially inconsistent use across studies (e.g., Bardovi-Harlig, 1992; Foster *et al.*, 2000).

Clausal complexity measures focus on clausal coordination and subordination. Again, t-units have played an important role for these measures because by its very definition, a t-

unit cannot contain more than one main clause. Thus, subordination can be measured in terms of the `number of clauses per t-unit` and coordination in terms of the `number of t-units per sentence`. (e.g., Kyle and Crossley, 2018; Lu, 2011). However, many other measures of clausal complexity have been proposed over the last decades that do not rely on t-units, both fine-grained (e.g., `number of relative clauses per clause`) and coarse-grained (e.g., `number of clauses per sentence`), see for example Graesser *et al.* (2004); Kyle and Crossley (2018); Vajjala and Meurers (2012). Global syntactic and clausal complexity measures have the longest tradition in absolute syntactic complexity assessment, whereas phrasal or sub-clausal complexity measures have been a more recent addition (Kuiken *et al.*, 2019; Kyle and Crossley, 2018; Norris and Ortega, 2009; Staples *et al.*, 2016). They focus on the modification and coordination of phrases with a special focus on noun phrases (NPs). Examples for phrasal complexity measures are for example `prenominal modifiers per NP` or `dependents per object`. Also measures outside of the nominal domain have been proposed, such as the `number of words preceding the main verb` (Graesser *et al.*, 2004). These types of measures have been shown to vary independently of clausal complexity across proficiency levels (e.g., Kyle and Crossley, 2018) and registers or tasks (e.g., Biber *et al.*, 2016; Staples *et al.*, 2016), see Section 2.1.3 for details.

All of the previously discussed measures have focused on syntactic elaboration. Considerably less work has been dedicated to syntactic variation (or 'syntactic diversity', see De Clercq and Housen, 2017). Some researchers have aggregated measures of clausal and phrasal structures not only using averages, but also with standard deviations to account for the variability of their occurrence (e.g., De Clercq and Housen, 2017; Kyle and Crossley, 2017). Other approaches to characterize the variability of syntactic structures have focused on counting the percentage of certain types of syntactic structures. For example, De Clercq and Housen (2017) propose to consider measures such as `number of matrix clauses per clause`, `number of subordinate clauses per clause`, and `number of coordinated clauses per clause` as estimates of clausal variation. However, it is debatable whether these are actually operationalizations of clausal *variation*. The same measures have been used in previous research as estimates of clausal *elaboration* (e.g., Hancke *et al.*, 2012; Vajjala and Meurers, 2012). Vercellotti (2019) tracks syntactic variety using the number of different clause types occurring in a language sample. Similarly, in Weiss (2015), I proposed to quantify phrasal variation in terms of the number of nominal modification types or deagentivation structures used. De Clercq and Housen (2017) proposed the `Syntactic Diversity Index`

to measure the diversity of clause types based on the logic of the MCI by Brezina and Pallotti (2019), see Section 2.1.2.4 for details.

Beyond this abundant work on absolute syntactic complexity, research has increasingly started to incorporate measures of relative syntactic complexity, for example in form of phraseological complexity (or 'syntactic sophistication', Kyle and Crossley, 2017) estimates. These target the use of collocations and frequent grammatical patterns. Researchers have focused on statistical collocations for specific grammatical relations (e.g., adjective-noun or verb-argument constructions) for a more targeted approach to phraseology (Kyle and Crossley, 2017; Paquot, 2019). This avoids, for example, considering frequent combinations of function words such as 'of the' (Paquot, 2018, p. 34). There has been ample research on these types of phraseological complexity measures for L2 writing and speech across different languages (e.g., Esfandiari and Ahmadi, 2022; Garner *et al.*, 2019; Hu *et al.*, 2022; Kyle *et al.*, 2021b; Paquot, 2018; Rubin, 2021; Vandeweerd *et al.*, 2021).

2.1.2.2 Lexical complexity

Lexical complexity, too, has a long tradition in complexity research (Bulté and Housen, 2012; Wolfe-Quintero *et al.*, 1998) and is the second most commonly studied domain of linguistic complexity (Kuiken *et al.*, 2019, p. 163). In contrast to the focus on elaboration in syntactic complexity, lexical complexity has predominantly been studied in terms of variation (Kuiken *et al.*, 2019, p. 163; Wolfe-Quintero *et al.*, 1998, p. 101). Also, the focus has been more evenly divided between measures of absolute and relative complexity in this domain. Measures of lexical complexity can broadly be categorized in two groups: text-internal measures and text-external measures (terminology from Skehan, 2009, p. 108). Examples for text-internal measures of absolute lexical complexity are lexical density (typically measured as the `number of lexical words per word`, Lu, 2012) and word length measures (e.g., `number of characters per word` or `number of syllables per word`) which have been particularly prevalent as global measures of lexical complexity in work on ARA (see Section 2.3). Most commonly, however, absolute lexical complexity is measured in terms of lexical diversity, sometimes also known as lexical variation (for an overview, see also Jarvis, 2013). Probably the most prominent measure of lexical diversity is the type token ratio (TTR) and its variants (e.g., Daller *et al.*, 2003, p. 199; Skehan, 2009, p. 108). The TTR compares the number of unique word form (types) to the total number of words (tokens), thus estimating how variable the vocabulary is that is used in a language sample. The measure has been used

in a wide range of research contexts such as language acquisition, SLA, forensic linguistics, stylometry, and clinical linguistics (for an overview, see Malvern *et al.*, 2004, p. 6–14). Some researchers have also proposed to measure part-of-speech (POS)-specific TTRs (e.g., Lu, 2012; Vajjala, 2018; Vajjala and Meurers, 2012).

The TTR has been shown to be strongly dependent on the number of words in a language sample because the repetition of word types becomes increasingly likely the longer a sample becomes (McCarthy and Jarvis, 2007, p. 460). To disentangle estimates of lexical diversity from sample length, several mathematical variations of the TTR have been proposed. Early suggestions include simple mathematical transformations such as the `root TTR`, the `bilogarithmic TTR`, or the `Uber index`, see Lu (2012) for details. However, these attempts had only limited success (see Malvern *et al.*, 2004; McCarthy and Jarvis, 2007). Alternatively, different sampling strategies have been proposed, such as the `D` measure in form of `vocd-D` (Malvern *et al.*, 2004) or the `Hypergeometric Distribution Diversity (HD-D)` by McCarthy and Jarvis (2007) as well as the `Measure of Textual Lexical Diversity (MTLD)` measure (McCarthy and Jarvis, 2010). Several studies have been proposed to test the validity of these different lexical diversity measures. Among the most influential was a series of studies conducted by McCarthy and Jarvis (2013, 2007, 2010). They compared a broad range of lexical diversity measures and found that MTLD was the most stable estimate (McCarthy and Jarvis, 2013, 2010), but that combining it with vocd-D or HD-D may help capturing additional dimensions of lexical diversity (McCarthy and Jarvis, 2010). Recently, some studies have attempted to link measures of lexical diversity to human judgments of lexical diversity in narrative writing (Jarvis, 2017) and argumentative writing (Kyle *et al.*, 2021a), finding medium to strong correlations between traditional holistic measures of lexical diversity (e.g., MTLD, HD-D). Jarvis (2017) and Kyle *et al.* (2021a) also suggest to measure lexical diversity as a multi-dimensional construct. They propose seven sub-dimensions including lexical volume (`number of word tokens`) and 'lexical abundance' (`number of lemma types`), while considering traditional holistic lexical diversity measures as measures of 'lexical variety'. However, also Kyle *et al.* (2021a) agree that traditional holistic lexical diversity measures are more suited for studies that take multiple complexity dimensions into consideration as they depend less on sample length.

Text-external measures rely on independent baseline values to estimate variation in the lexical domain. Most text-external measures assess relative lexical complexity. The most prominent example for this are word frequency estimates. More frequent words are acquired earlier

in L2 acquisition (Ellis, 2002) and are processed and retrieved faster (for an overview, see Brysbaert *et al.*, 2011, 2018), although there is evidence suggesting that the effect weakens with increased exposure to a language (Brysbaert *et al.*, 2017; Diependaele *et al.*, 2013). They have been commonly used in work on language reception, including ARA, see Collins-Thompson (2014) for an overview. They have also been shown to be valuable estimates of language performance in L1 and L2 contexts (Crossley, 2020; Crossley *et al.*, 2012; De Jong, 2016; Ellis, 2002; Kim *et al.*, 2018; Vermeer, 2001). More proficient speakers are generally assumed to have access to a larger vocabulary, which allows them to utilize infrequent vocabulary (Bulté and Housen, 2014, p. 50; Crossley, 2020, p. 418), at least when it is functionally adequate, e.g., in academic language use (see discussion of adequacy in Section 2.1.3.2, p. 37).

Frequency measures rely on external data bases of word frequencies that were obtained from large language samples (about 20 million words, cf. Brysbaert *et al.*, 2011, 2018). These data bases not only need to be sufficiently large, but also representative for the language experience of the target group. Recent studies suggest that it is beneficial to tailor the frequency norm to match the specific language experience of the target group rather than using traditional frequency data bases that were compiled from newspapers and books several decades ago (Brysbaert *et al.*, 2018, p. 45). Several studies have shown that frequencies from television or social media data are better suited for students (e.g., Brysbaert and New, 2009; Dimitropoulou *et al.*, 2010; Gimenes and New, 2016) than traditional frequency data bases based on written language. In contrast, there is some evidence suggesting that these traditional sources are better suited to model the language experience of older target groups (Cuetos *et al.*, 2012a). Frequency measures are commonly measured as average (log) frequencies.[1] Alternatively, frequencies have been expressed in terms of log frequency bands (e.g., Hancke, 2013). More recently, Chen and Meurers (2018) have proposed to use standard deviations of word frequencies as additional frequency measures to capture more information of the distribution of word frequencies in a language sample.

Other commonly used text-external measures of relative lexical complexity are based on word lists, age of acquisition (AoA), and contextual diversity. Word lists of both, basic vocabulary (e.g., Ács *et al.*, 2013; Chiari and De Mauro, 2014; Thorndike, 1921) and academic vocabulary (e.g., Coxhead, 2000; Gardner and Davies, 2014) have a long tradition in research on ARA (see overviews by Collins-Thompson, 2014; DuBay, 2004, 2006) and studies on writ-

[1] For example: `sum of frequencies of words from the language sample that were found in data base X divided by the number of words in the language sample that were found in data base X.`

ing quality and development (e.g., Bestgen, 2017; Laufer and Nation, 1995; Olinghouse and Leaird, 2009; Yoon, 2018). They are often frequency-based and used as more coarse-grained estimates of how much simple or sophisticated vocabulary a language sample contains. To account for discipline specific variation of academic language use (e.g., Durrant, 2016), several specialized academic word lists have been proposed, for example for medicine (Wang *et al.*, 2008), environmental science (Liu and Han, 2015), or finance (Li and Qian, 2010). Also AoA estimates have been shown to cover a relevant dimension of lexical processing. Rather than focusing on frequency, they quantify the age at which L1 speakers acquire words. AoA estimates have traditionally been obtained through subjective ratings by adult native speakers, but there has also been some work on obtaining more objective AoA estimates through picture naming experiments with children (for an overview, see Bonin *et al.*, 2004, pp. 457–458). Despite correlations with frequency measures, AoA has shown to have a relevant effect on lexical processing in its own right in several empirical studies (e.g., Brysbaert and Cortese, 2011; Brysbaert and Ghyselinck, 2006; Cuetos *et al.*, 2006). Finally, rather than only counting the frequency of words, researchers have proposed to also consider the number of contexts in which words appear. This has been referred to as contextual diversity. Several studies have shown the relevance of this factor for vocabulary learning and word recognition (e.g., Adelman *et al.*, 2006; Hills *et al.*, 2010; Johns *et al.*, 2016).

2.1.2.3 Semantic complexity

Relatively little work has focused on semantic complexity, but some computational linguistic approaches to text assessment have included measures of semantic elaboration, variation and relatedness (e.g., Brück and Hartrumpf, 2007a; Crossley *et al.*, 2010a; Hancke, 2013; Kim *et al.*, 2018; Venant and d'Aquin, 2019; vor der Brück *et al.*, 2008). Work exploring semantic elaboration, variation and relatedness has focused on quantifying hierarchical (hyponymy, hypernymy) and lateral (synonymy, antonymy, polysemy) semantic relations and semantic fields using word nets and knowledge graphs as external resources. Also measures of negations and word concreteness and imagability have been proposed to assess relative semantic complexity (e.g., Brück and Hartrumpf, 2007b; Crossley *et al.*, 2011b). Some work on word processing, too, has focused on the semantic diversity (e.g., Hoffman *et al.*, 2013) and distinctiveness (e.g., Johns and Jones, 2022; Johns *et al.*, 2012) of words based on their distributional properties. Generally, Latent Semantic Analysis (LSA) and word embeddings have been used systematically in computational linguistics to tap into the semantic dimension of language

productions and evaluate their quality (e.g., Briscoe *et al.*, 2010; Crossley and McNamara, 2011; Crossley *et al.*, 2011b) or accessibility (e.g., Hancke *et al.*, 2012; Hsiao and Nation, 2018). However, these measures are often used independent of the notion of 'complexity'.

It should be noted that semantic complexity is usually not identified as an independent dimension of complexity. It is, for example, absent in the taxonomies proposed by Bulté and Housen (2012) and Housen *et al.* (2019). Instead, measures of semantic complexity are commonly subsumed under the notion of lexical (or lexico-semantic) complexity—as 'lexical relatedness'—(e.g., Crossley, 2020; Hancke, 2013; Pallotti, 2015a). However, in view of the multitude of sub-dimensions of lexical complexity, it might be useful to discuss lexical and semantic complexity separately, especially as measures of semantic complexity seem to be less predictive of human judgments of lexical proficiency than, for example, lexical diversity measures (Crossley *et al.*, 2011b). Also, some measures have been referred to as estimates of semantic complexity but are at the same time commonly associated with discourse complexity, e.g., the the use of relational chains (causal, concessive, temporal), the `number of propositions`, or the use of anaphors (e.g. Brück and Hartrumpf, 2007a; Brück *et al.*, 2008). A clearer conceptual distinction between the notions of lexical, semantic, and discourse complexity is needed. Considering semantic complexity as a dimension in its own right might encourage more conceptual clarity while also fostering methodological innovations regarding the assessment of semantic complexity (see for example Venant and d'Aquin, 2019).

2.1.2.4 Morphological complexity

Morphological complexity has been studied considerably less intensively in SLA complexity research than syntactic or lexical complexity (see also Housen *et al.*, 2019, p. 12; Kuiken *et al.*, 2019, p. 163). There has been relatively little work on morphological complexity in research on early First Language Acquisition (FLA) (Xanthos and Gillis, 2010, p. 176). Bulté (2013, p. 88) suggests that this lack of work on morphological complexity was facilitated by the strong focus of complexity research on English. This parallels ARA research, which has successfully used morphological measures for non-English languages (e.g. Dell'Orletta *et al.*, 2011; François and Fairon, 2012; Hancke *et al.*, 2012; Reynolds, 2016), but rarely for English. Yet, morphological complexity lends itself to cross-linguistic comparisons and the study of complexity trade-offs across linguistic domains. Therefore, typological research has paid great attention to morphological complexity, see for example Baerman *et al.* (2015) and contributions therein as well as several contributions in Miestamo *et al.* (2008).

More work on morphological complexity has started to appear in studies on L1 development (e.g., Malvern *et al.*, 2004; Xanthos and Gillis, 2010) and L2 development (e.g., Brezina and Pallotti, 2019; Bulté, 2013; De Clercq and Housen, 2019; van der Slik *et al.*, 2019). Absolute morphological complexity has predominantly been researched in terms of inflectional diversity, for example in form of the `Inflectional Diversity` measure by Malvern *et al.* (2004), the `Normalized Mean Size of Paradigm` by Xanthos and Gillis (2010), and, more recently, the `Morphological Complexity Index (MCI)` by Pallotti (2015b), see Brezina and Pallotti (2019) for detailed discussion in English. The first two measures calculate morphological diversity independent of POS and rely on the full inflected word forms. In contrast, Pallotti's MCI compares the variability of morphological exponents for specific POS (e.g., verbs, nouns, adjectives). Despite this difference, all three measures are closely related to measures of lexical diversity (see Section 2.1.2.2) and are prone to the same length dependency issue. To mitigate their dependency on length, all three measures utilize different randomized re-sampling strategies (for details, see Brezina and Pallotti, 2019; Malvern *et al.*, 2004; Xanthos and Gillis, 2010). As far as I am aware, there has been no systematic evaluation of the empirical validity of these measures comparable to McCarthy and Jarvis's (2010) work on lexical diversity. However, De Clercq and Housen (2019) compare three measures of morphological diversity for the assessment of L2 proficiency in French and English speech produced by adolescents: `Inflectional Diversity`, MCI, and a simple `word type to word family ratio` that has been used by Horst and Collins (2006) to measure morphological diversity in adolescents' English L2 writing. Concluding their comparison, De Clercq and Housen (2019) recommend using the MCI to assess morphological diversity because it computes morphological diversity based on morphological exponents rather than inflected word forms (De Clercq and Housen, 2019, p. 92). This makes it less sensitive to effects from differences in lexical diversity and more suited for shorter language samples (see also discussion in Brezina and Pallotti, 2019, p. 101–103).

Little research has focused on local measures of morphological elaboration. However, there has been some work on German focusing on the elaboration of compound nouns, for example calculating the `number of compounds per compound noun` or `percentage of compound nouns` (Brück *et al.*, 2008; Hancke *et al.*, 2012). Some measures have been proposed to track the presence of specific derivational processes (such as `percentage of derived nouns`) or inflectional processes linked to tense, mood, aspect, and case (such as `percentage of verbs in simple past` or `percentage of nouns with genitive case marking`).

These measures have been particularly prevalent in work on ARA for languages other than English (e.g., François and Fairon, 2012; Hancke *et al.*, 2012; Reynolds, 2016) but also in SLA work on L2 proficiency (e.g., Guo *et al.*, 2013; Verspoor *et al.*, 2012). These measures are typically not motivated from a language system perspective. For example, simple past is not argued to be a more elaborate tense than simple present. However, they have been identified as challenging or acquired at later stages in L2 or L1 development. In that sense, such measures can be considered estimates of relative morphological complexity (see also Bulté, 2013, p. 89; De Clercq and Housen, 2019, p. 75).

2.1.2.5 Phonological complexity

Phonological complexity has been studied extensively for cross-linguistic comparisons in linguistic typology (e.g., Pellegrino *et al.*, 2009, and contributions therein). In this context, phonological complexity has been predominantly studied in terms of a language's inventory of consonants, vowel contrasts and qualities, and tonal systems or in terms of the complexity of their syllables (Maddieson, 2009). Some work has also attempted to combine these dimensions of phonological complexity into holistic scores, e.g., Atkinson's (2011) measure of `phoneme diversity`, which calculates the language-wise average score of three normalized sub-dimensions of phonological complexity: size of the consonant inventory, size of inventory of basic vowel qualities, and complexity of the tone system. Maddieson *et al.* (2011) confirm the empirical validity of this measure even though they argue that conceptually it fails to globally quantify the diversity of a language's phonological inventory (for a more detailed discussion, see Maddieson *et al.*, 2011, p. 268). Using the terminology from Section 2.1.1, these types of measures assess the phonological elaborateness and variation of a language in contrast to another language. Maddieson (2009) also discusses work on relative phonological complexity in terms of frequency measures for phonological units (e.g., segments).

Also psycho-linguistic research has considered aspects of phonological complexity (e.g., in terms of `word length in syllables`) and phonological (dis)similarity, (e.g., using the `phonological similarity metric analysis` by Mueller *et al.*, 2003). Research has mostly focused on understanding the effects of phonological complexity on aspects such as verbal working memory and speech planning times (for an overview, see Mueller *et al.*, 2003). However, there has also been some work on linking phonological complexity to language learnability, see Gierut (2007) for a review. In SLA complexity research, however, phonological complexity has so far played virtually no role (Housen *et al.*, 2019, p. 11; Kuiken *et al.*, 2019,

p. 163).

2.1.2.6 Discourse complexity

SLA complexity research has put little emphasis on the dimension of relative discourse complexity in terms of cohesion. Discourse cohesion (sometimes also called 'text cohesion') refers to the use of linguistic devices to connect meaning units in language productions (e.g., Crossley, 2020; Todirascu *et al.*, 2016). These linguistic devices can make discourse relations explicit (e.g., in form of connectives) or create implicit links (e.g., anaphors and repetitions). Cohesive devices can be used to connect ideas between adjacent clauses (local cohesion) or across larger organizational units such as paragraphs (global cohesion). Text cohesion is a dimension of relative discourse complexity. FLA and writing research has demonstrated that global and local cohesive devices are adopted relatively late into L1 speakers' language production (see overview in Crossley, 2020, p. 427). Also, research on FLA has demonstrated processing difficulties related to the resolution of anaphors and other implicit cohesion devices in young L1 speakers (e.g., Englert and Hiebert, 1984; Gülzow and Gagarina, 2007; Joseph *et al.*, 2015; Yuill and Oakhill, 1988).

The use of cohesive devices helps listeners or readers to create and maintain a mental representation of the discourse and correctly connect all meaning units in it—it promotes 'coherence' (Graesser *et al.*, 2003; Louwerse *et al.*, 2004). This has made cohesion an important dimension in ARA (e.g., Collins-Thompson, 2014; McNamara and Graesser, 2012; McNamara *et al.*, 2010a; Todirascu *et al.*, 2016). From a language production perspective, the appropriate use of cohesive devices to organize ideas and communicate them clearly has been viewed as an indicator of language proficiency and writing quality (e.g., Crossley, 2020; Crossley and McNamara, 2016a; Crossley *et al.*, 2016c). In both application domains, measures of text cohesion have mostly focused on four types of cohesive devices: connectives, co-reference, the repetition of lexical items, and discourse entities (Collins-Thompson, 2014; Crossley, 2020). Explicit cohesion has been measured predominantly in terms of the use of specific connectives, e.g., `causal connectives per clause` or `temporal connectives per 1.000 words` (Berendes *et al.*, 2018; Crossley *et al.*, 2016c; McNamara and Graesser, 2012). While the use of adverbial clauses that are introduced with specific connectives itself results in greater clausal elaboration, the focus of these measures has been on the types of discourse relations that they introduce into the language sample (e.g., Crossley, 2020; McNamara and Graesser, 2012; Myhill, 2008). However, there is an undeniable correlation between

syntactic elaboration and these measures of text cohesion that makes the clear separation of measures for these two dimensions at times challenging (Crossley, 2020, p. 417).

Implicit cohesion has been assessed through a broad variety of measures. To assess co-reference, researchers have proposed simple POS-based measures of 'givenness' that track the use of pronouns, determiners, and proper nouns (e.g., Crossley *et al.*, 2016c; Pitler and Nenkova, 2008; Todirascu *et al.*, 2013). Examples for such measures are `pronoun to noun ratio` or `definite articles per sentence`. Also co-reference chains have been used to quantify implicit cohesion (e.g., Todirascu *et al.*, 2013; Xia *et al.*, 2016). Furthermore, implicit cohesion has been measured through the density of discourse entities, for example in form of `number of discourse entities` (per sample or sentence) or the `number of named entities` (Feng *et al.*, 2009, 2010; Todirascu *et al.*, 2013). Brown *et al.* (2008) proposed the propositional idea density (PID) measure to track global implicit cohesion. The measure quantifies the number of propositions in a language sample using the notion of propositions developed by Kintsch (1974) and Turner and Greene (1977). Propositions have shown to play a relevant role for retention and memory (Kintsch and Keenan, 1973; Perrig and Kintsch, 1985) and PID has been successfully used to predict cognitive impairments (Bryant *et al.*, 2013; Jarrold *et al.*, 2010; Sirts *et al.*, 2017). A third way to quantify implicit cohesion is through measures of lexical repetition across sentences and paragraphs (e.g., Crossley *et al.*, 2016c; McNamara *et al.*, 2010a; Vajjala, 2018). For example, Coh-Metrix calculates overlaps of nouns, arguments, content words, and stems locally (between adjacent sentences) or globally (between all sentences), see for example McNamara *et al.* (2010a). Crossley *et al.* (2016c) additionally propose to calculate the semantic overlap across synonyms, hyponyms, and hypernyms. Barzilay and Lapata (2008) proposed to track the occurrence of discourse entities in different grammatical roles (subject, object, other, absence). To do so, they build a two-dimensional entity grid (entities $\times$ sentences) across all sentences in a language sample. This can be used to identify common transition patterns or calculate transition probabilities between grammatical roles, both locally (i.e.,between adjacent sentences) and globally (across all sentences). For example, the `local transition probability from subject to object role` provides the probability of a discourse entity to occur as a subject at sentence i and as object in sentence $i+1$. Several studies have used these entity-grid-based measures to quantify cohesion (e.g, Pitler and Nenkova, 2008; Todirascu *et al.*, 2013; Vajjala, 2018; Weiss and Meurers, 2018; Xia *et al.*, 2016).

Some research promotes the use of LSA to measure the semantic similarity between para-

graphs as a measure of global implicit cohesion (e.g., Crossley, 2020; Foltz, 2007; McNamara *et al.*, 2007). The underlying idea is that greater semantic similarity indicates a more cohesive discourse. However, these measures interact with aspects of semantic complexity because greater semantic similarity also indicates lower semantic variation. While it is intuitive that high semantic complexity can impede cohesion, this interaction requires further investigation. This emphasizes the need for a stronger methodological and conceptual reappraisal of the concepts of semantic complexity, cohesion, and their interaction.

2.1.2.7 Processing complexity

Processing complexity is a dimension that I propose to group established complexity measures that focus on quantifying costs of information processing and do not pertain to any of the previous dimensions. These information processing costs may occur in computational or cognitive processing. The most prominent types of measures from this dimension are information-theoretic measures of global absolute complexity. In information theory, 'information' is commonly understood in terms of its unexpectedness or 'entropy', following Shannon (1948). The so called 'Shannon entropy' has been used to quantify the information in a variable or message in bits. For the present discussion, it suffices to know that following this notion, a high entropy (i.e., surprising) variable or message carries more information than a low entropy (i.e., unsurprising) variable or message. Thus, higher entropy equals higher complexity (for a more detailed discussion, see Juola, 2008, pp. 90–91). Another information theoretic approach to complexity is the Kolmogorov complexity (Li and Vitányi, 2008). It measures the complexity of a string in terms of the length of its shortest possible description. A string that can be described concisely is less complex than a string requiring a longer explanation. Compression-algorithm-based measures have been proposed as operatiolanizations to approximate the Kolmogorov complexity (Juola, 2008, pp. 91–93). The intuition behind this is that compression algorithms shrink the bits required to represent a string by eliminating its redundancies. In a way, they attempt to find the most efficient characterization of that string. Low entropy strings containing many redundancies can be compressed more than high entropy strings containing few to no redundancies. Compression-based complexity measures have been particularly popular as operationalizations of global, absolute complexity for cross-linguistic studies (e.g., Ehret, 2018; Ehret and Szmrecsanyi, 2016; Juola, 2008). Recently, they have been used in SLA complexity research as well to assess the link between complexity and L2 essay quality (Ehret and Szmrecsanyi, 2019).

Human information processing costs have been studied in psycho-linguistic and computational linguistic research from different perspectives (for an overview, see Levy, 2013). Measures derived from this research can be understood as estimates of relative processing complexity, because most research has focused on explaining empirically observed processing difficulties of humans, for example through eye-tracking or processing time estimates. There are two main strands of research on incremental human sentence processing and processing difficulties, which Futrell *et al.* (2021) refer to as 'expectation-based theories' and 'memory-based theories'. In expectation-based theories, information theoretic notions of entropy or surprisal have played a central role. Measures derived from these theories operationalize complexity in terms of word probabilities given their context. These word probabilities have been obtained through probabalistic context-free grammars (PCFGs; e.g., Demberg and Keller, 2008; van Schijndel *et al.*, 2013) or various types of probabilistic language models (e.g., Goodkind and Bicknell, 2018; Hao *et al.*, 2020; Monsalve *et al.*, 2012). These types of surprisal measures have been used successfully to model processing times and eye-tracking data as well as locality effects in human sentence processing (e.g., Futrell and Levy, 2017; Monsalve *et al.*, 2012; Oh *et al.*, 2021).

Memory-based theories of human sentence processing hypothesize that the integration of new information into the mental representation of a listener or reader requires the retrieval of previously processed words from working memory. This retrieval can be more or less cognitively taxing and may cause processing difficulties in certain linguistic contexts. A prominent example of such a memory-based theory is the Dependency Locality Theory (DLT) by Gibson (2000). The theory assumes that incremental sentence processing requires working memory for two components: storing incomplete discourse structures and integrating new discourse referents into incomplete discourse structures which have to be retrieved from memory. Additionally, the DLT supposes that the costs of integrating a discourse referent into an incomplete discourse structure increases with the number of intervening discourse referents. This accounts for the so called 'locality effect', which has shown to play a role for sentence processing in several studies (e.g., Bartek *et al.*, 2011; Fedorenko *et al.*, 2013; Grodner and Gibson, 2005; Nicenboim *et al.*, 2015). Shain *et al.* (2016) proposed different configurations of Gibson's original cost calculation algorithm to account for higher verb weights and reducing the processing costs for coordinated phrases and modifiers. These are procedural measures of relative processing complexity. In Weiss (2017), I proposed a method to aggregate the processing costs of sentences at the point of their maximum cost across all sentences in a language

sample to derive a single measure. To do so, I proposed measures of maximum and average processing costs (for details, see Weiss, 2017, pp. 73–75). Other computational linguistic measures of memory-based processing costs are based, for example, on left-corner parsers (e.g., Shain *et al.*, 2016; van Schijndel *et al.*, 2013), which have been used for decades to quantify human processing costs (Resnik, 1992). There have been some attempts to combine the expectation-based and the memory-based perspective on incremental human sentence processing. However, these unified models require more work to account for interaction effects between high surprisal and memory demands, see Futrell *et al.* (2021, pp. 8–9) for a detailed discussion.

2.1.3 Variation in linguistic complexity

Variation is a central construct in applied linguistics. It has been studied in functional and usage-based approaches across linguistic subdisciplines, especially in sociolinguistics (e.g., Eckert, 2012; Tagliamonte, 2006) and linguistic typology (e.g., Cacoullos and Travis, 2019; Sinnemäki, 2020). Speaker preferences between alternative surface forms have also been studied in generative and system-based approaches, see Adli *et al.* (2015) for a general comparison between these approaches. In SLA research, linguistic variation factors into the characterization of L2 performance (Kuiken *et al.*, 2019). Learners produce language by choosing from a pool of more or less adequate options and may exhibit intra-individual variation in their choices or inter-individual variation across groups of learners. These choices can be influenced by a broad range of factors including but not limited to the linguistic forms available in the (inter-)language system, the production context and modality, individual differences, stylistic choices, and differences between languages. In the following, I zoom in on three important dimensions of variation in linguistic complexity that are immediately relevant for this thesis: language development and proficiency, task and register, and cross-lingual variation.

2.1.3.1 Developmental variation

Language development and proficiency differences are probably the best studied sources of variation in the expression of CAF (sub-)dimensions. In SLA complexity research, complexity has been predominantly used "(a) to gauge proficiency, (b) to describe performance, and (c) to benchmark development" (Ortega, 2012, p. 128). Even though 'development' and 'proficiency' are independent constructs (for details, see Section 2.2.1 or Bulté and Housen,

2014), proficiency is often used to approximate development on cross-sectional data (Ortega and Iberri-Shea, 2005, p. 27). This is motivated by the fact that L2 development is assumed to lead to an extension of learners' inter-language system and a stronger approximation to the target norm (i.e., higher proficiency, Bulté and Housen, 2014, p. 46)—at least in instructed settings. For the remainder of this section, I will discuss aspects of development and proficiency jointly.

Previous research successfully demonstrated that the linguistic complexity of learner productions systematically varies with proficiency across linguistic domains (Housen *et al.*, 2012, 2019; Kuiken *et al.*, 2019; Norris and Ortega, 2009; Skehan, 2009; Wolfe-Quintero *et al.*, 1998). This research has largely focused on the domains of syntax and lexicon followed by morphology and text cohesion. In the following, I focus my discussion on these domains.

Syntax For the analysis of syntax, most research has focused on written academic language development. In this context, studies on L1 writing in school and university have repeatedly found evidence that phrasal complexity increases in more advanced writing while clausal complexity decreases (e.g., Crossley *et al.*, 2011a; McNamara *et al.*, 2010b; Staples *et al.*, 2016). The role of register on this observation is discussed in Section 2.1.3.2 (p. 36). A similar development has been reported for L2 writing, which typically focused on learners at the college level or university level but partially also included studies in schools (e.g., Bulté and Housen, 2018, 2019; De Clercq and Housen, 2017). Several studies have linked higher phrasal complexity—especially in the nominal domain—to writing of more advanced or proficient learners (e.g., Bulté and Housen, 2018, 2019; Lu, 2011; Yoon, 2017; Yoon and Polio, 2017) and to higher ratings of writing quality (e.g., Guo *et al.*, 2013; Kyle and Crossley, 2017, 2018; Taguchi *et al.*, 2013). In contrast, developmental variation within clausal domain has been described with mixed findings. Several studies have reported that in instructed settings, English L2 learners develop increasingly elaborate clausal structures through subordination (e.g., Alexopoulou *et al.*, 2017; Bulté and Housen, 2018, 2019; Lu, 2011; Norris and Ortega, 2009; Verspoor *et al.*, 2012) and that for advanced learners clausal elaboration stagnates or even reduces in a trade-off with phrasal elaboration (Bulté, 2013; Kyle and Crossley, 2018; Lu, 2011; Norris and Ortega, 2009; Taguchi *et al.*, 2013; Yoon, 2017; Yoon and Polio, 2017). Others found that even though subordination was positively correlated with writing quality ratings, the clausal domain of writings by intermediate to advanced adult English as a Second Language (ESL) learners did not develop in terms of subordination (Bulté and Housen,

2014; Crossley and McNamara, 2014), but in terms of coordination (Bulté and Housen, 2014). Bulté and Housen (2014) conclude from this that advanced academic L2 writing is somewhat flexible in the development of the syntactic domain. It also highlights the need to distinguish between writing quality ratings and development (Crossley and McNamara, 2014; Crossley *et al.*, 2014a).

Studies on developmental variation in spoken language are less common. De Clercq and Housen (2017) found that advanced adolescent French and English L2 learners' oral narratives were globally more elaborate and varied than less advanced learners' narratives, but did not form more elaborate or varied NPs. This is in line with Biber *et al.*'s (2011) claim that phrasal complexity is a characteristic of academic written language, whereas spoken language is characterize by higher clausal complexity. Similarly, Lambert and Nakamura (2019) analyzed the syntactic complexity of intermediate English L2 speakers' descriptive speech. They found that clausal elaboration increased with proficiency. For phrasal elaboration, they found a considerable interaction with learners' knowledge of task-relevant vocabulary. In contrast, Vercellotti (2019) found a simultaneous increase of clausal and phrasal complexity in spoken monologues of low-intermediate to high-intermediate English L2 learners in the course of one semester of instruction. While these findings seem to confirm the generally assumed tendency of spoken language to develop in terms of clausal elaboration, more research is needed to understand the developmental variation of phrasal complexity in L2 speech.

Lexicon Findings regarding the developmental variation of sub-dimensions of lexical and semantic complexity have been more straightforward. Several studies showed that higher lexical diversity in L2 writing is linked to longitudinal development in instructed settings and higher writing quality ratings (e.g., Alexopoulou *et al.*, 2017; Bulté and Housen, 2019; Crossley and McNamara, 2012; Crossley *et al.*, 2010b; Guo *et al.*, 2013; Yoon, 2017; Zheng, 2016). This finding has also been confirmed for French and Italian L2 writing (Kuiken and Vedder, 2012). Studies also showed that more advanced L2 writing used more sophisticated vocabulary in terms of word frequencies (Crossley and McNamara, 2012; Guo *et al.*, 2013; Jung *et al.*, 2019; McNamara *et al.*, 2010b; Yoon, 2017; Zheng, 2016) and—albeit studied less often—word familiarity, concreteness, associations, precision or specificity, and AoA (e.g., Crossley and McNamara, 2012; Crossley *et al.*, 2010b; Guo *et al.*, 2013; Jung *et al.*, 2019; Kim *et al.*, 2018; Kyle and Crossley, 2016). Crossley *et al.* (2011a) found that lexical sophistication (measured through word frequency and concreteness) also increased for adolescents' English

L1 writing across grade levels. Crossley *et al.* (2010b) evaluated the link between lexical complexity and essay quality ratings for both, English L1 and L2 writings. In line with the previous findings reported for L2 writing and L1 development, they found that higher rated essays were lexically more diverse and polysemous, contained less concrete, familiar and imaginable vocabulary, and used longer words.

For L2 speech, studies found that lexical variation increases with proficiency (Crossley *et al.*, 2009, 2011b; Horst and Collins, 2006) and is positively correlated with human judgments (Lu, 2012). However, vocabulary in English L2 speech also seems to become more polysemous (Crossley *et al.*, 2010a, 2011b), more frequent (Crossley *et al.*, 2010a), and more variable in its frequency (Horst and Collins, 2006). It is also unclear whether L2 speech becomes more specific (Crossley *et al.*, 2009) or less specific (Crossley *et al.*, 2011b). Furthermore, Lu (2012) reported that lexical sophistication and density do not play a relevant role in human quality estimates of English L2 oral narratives.

Morphology As discussed in Section 2.1.2.4, SLA complexity research on the morphological domain has been considerably more limited than work on syntactic and lexical complexity. However, in recent years, several studies have investigated developmental variation in the morphological domain of L2 learners. Research has nearly exclusively focused on the inflectional diversity, predominantly in the verbal domain (Brezina and Pallotti, 2019; Bulté and Housen, 2019; De Clercq and Housen, 2019; Xanthos *et al.*, 2011), but partially also for other POS (Xanthos and Gillis, 2010; Xanthos *et al.*, 2011) or POS-independent (De Clercq and Housen, 2019; Malvern *et al.*, 2004). For L2 writing, the studies systematically showed that for beginning to high-intermediate learners of English, French, and Italian, morphological variation in the verbal domain increases with higher proficiency (Brezina and Pallotti, 2019; Bulté and Housen, 2019; De Clercq and Housen, 2019; Yoon, 2017). However, Brezina and Pallotti (2019) report that for both, L2 Italian and L2 English, the developmental variation of morphological complexity seems to level off at a high-intermediate to advanced proficiency level (Brezina and Pallotti, 2019)—albeit earlier for English than for Italian. De Clercq and Housen (2019) made similar observations for oral narratives of adolescent L2 speakers of French and English at beginning to advanced proficiency levels. They report that for both languages, learners' speech became morphologically more diverse but that this development leveled off at higher proficiency levels.

For early FLA until the age of three, the inflectional variation of children's speech has been

shown to increase for English and Spanish (Malvern *et al.*, 2004), Dutch (Xanthos and Gillis, 2010), as well as for French, Dutch, German, Russian, Croatian, Greek, Turkish, Finnish and Yucatec Maya (Xanthos *et al.*, 2011). Xanthos *et al.* (2011) report that the speed of the development was modulated by the morphological variation in the language that caregivers directed at children. Furthermore, Xanthos and Gillis (2010) and Xanthos *et al.* (2011) found that across languages, verb inflection showed stronger developmental variation than noun inflection with Xanthos and Gillis (2010) being unable to confirm any development for Dutch L1 learners between age two and three.

Text cohesion Developmental variation in text cohesion has been studied in the context of research on essay quality of high school, college, or university students with English as their L2 (e.g., Crossley and McNamara, 2012; Guo *et al.*, 2013; Jin, 2001; Jung *et al.*, 2019) or L1 (e.g., Crossley and McNamara, 2016b; Crossley *et al.*, 2014b, 2016c; McNamara *et al.*, 2010b). Work on elementary school students has focused on L1 expository and narrative writing (e.g., Cain, 2003; Cameron *et al.*, 1995; Cox *et al.*, 1990; Struthers *et al.*, 2013). For intermediate to advanced L1 and L2 writing, global cohesion has been shown to increase with proficiency (Crossley and McNamara, 2009, 2016b; Crossley *et al.*, 2016c; Jung *et al.*, 2019). For local cohesion, mixed empirical results have been reported. Crossley and McNamara (2009) found adults' L1 writing uses systematically more explicit and implicit local cohesion devices than advanced L2 writing. Other studies found that L2 essays with higher local cohesion received higher proficiency ratings (Crossley and McNamara, 2016b; Guo *et al.*, 2013) and that elementary school children's L1 writing was rated higher when it used more local cohesive devices (Cain, 2003; Cameron *et al.*, 1995; Cox *et al.*, 1990; Jin, 2001). However, even more studies reported no effect—or even a negative effect—of local cohesion on expert proficiency ratings. This was shown for children (Fitzgerald and Spiegel, 1986; Spiegel and Fitzgerald, 1990) as well as for adults' L1 essays (Crossley and McNamara, 2012; Crossley *et al.*, 2016c; McNamara *et al.*, 2010b) and L2 essays (Guo *et al.*, 2013; Jung *et al.*, 2019). Crossley *et al.* (2011a) reported a reduction of local cohesion in adolescents' essays across grade levels. This is at odds with the assumption that higher textual cohesion fosters coherence seeing that coherence is a strong predictor of text quality (e.g., Crossley and McNamara, 2010, p. 988). It also contrasts research on text readability which has found that higher local cohesion fosters readability (e.g., Best *et al.*, 2006; Collins-Thompson, 2014).

Several explanations have been proposed to reconcile these contradictory findings. First, it

has been argued that the 'reverse cohesion effect' might cause the negative correlation between expert ratings and local cohesion (Crossley and McNamara, 2012; Crossley *et al.*, 2011a; Guo *et al.*, 2013; McNamara *et al.*, 2011). The reverse cohesion effect describes the phenomenon that readers with high prior knowledge on the text topic can benefit from a text because the need to infer connections based on their prior knowledge promotes their cognitive activation (Crossley and McNamara, 2010; McNamara, 2001, for details on prior knowledge and reading comprehension, see Section 2.3.1.1). In contrast, readers with less prior knowledge struggle to make these inferences on their own and require a text with more cohesive devices (Best *et al.*, 2008; Crossley and McNamara, 2010; McNamara, 2001). This aligns with Crossley and McNamara's (2010) finding that less locally cohesive L1 essays are rated as more coherent by human experts. However, other studies found that the reverse cohesion effect impacts high-knowledge readers with low reading comprehension skills and is irrelevant for readers with high reading skills (O'Reilly and McNamara, 2007; Ozuru *et al.*, 2009). This speaks against the reverse cohesion effect as an explanation for the negative impact of local cohesion on essay ratings, because expert raters can be assumed to be highly skilled readers. Second, Guo *et al.*'s (2013) study showed that task context seems to interact with human ratings and text cohesion. While they reported a negative effect of local cohesion on essay quality for independent writing, they found that higher local cohesion was positively correlated with writing quality in integrated writing, even though lexical sophistication remained a stronger predictor. Finally, differences in writing strategies may explain some of the mixed results. Crossley *et al.* (2014b) studied profiles of highly rated L1 essays written by high school and college students. Their cluster analysis identified four different writing styles that differ considerably in the linguistic means that they employ: a) action and depiction, b) academic style, c) accessible style, and d) lexical style. All styles but the accessible style are characterized by higher lexical complexity and are either agnostic to cohesion or exhibit lower textual cohesion. In contrast, high local cohesion and less sophisticated vocabulary were characteristic for the accessible writing style.

These explanations shed some light on the mixed findings regarding the developmental variation of local cohesion. Three other important factors that need to be investigated further are genre and language. As noted earlier, much research on the developmental variation of textual cohesion has focused on English high-intermediate to advanced academic essay writing. It is unclear which role cohesion plays for other text genres and especially non-academic language use. The findings might also not transfer to other languages in which the academic language register is characterized by different linguistic structures (see Section 2.1.3.2 and

Section 2.1.3.3 for a more detailed discussion). In short, despite its prominence in writing quality research and readability research, more studies are needed to understand the developmental variation of cohesion in FLA and SLA.

The overview provided in this section demonstrates that linguistic domains vary independently with development and proficiency. The discussion of differences between developmental variation in beginning and advanced learners has also highlighted a commonly stated fact about complexity development. Complexity is not necessarily increasing linearly with proficiency (see e.g., Larsen-Freeman, 2009; Pallotti, 2009). The development of a certain linguistic domain may level off, temporarily stagnate or—in the case of u-shaped developmental patterns (Ellis and Larsen-Freeman, 2006, p. 595; VanPatten and Benati, 2010, p. 141)—regress. Researchers have also reported trade-off effects between linguistic domains and subdimensions of complexity, for example between phrasal and clausal elaboration (e.g., Bulté and Housen, 2018; Lambert and Nakamura, 2019; Lu, 2011; Ortega, 2003). Finally, developmental variation can become non-linear through interactions with task or register variation and may differ across languages. These two factors will be discussed in Section 2.1.3.2 and Section 2.1.3.3.

2.1.3.2 Task and register variation

Task factors and register have been shown to influence the expression of linguistic complexity in language productions of non-native speakers (e.g., Alexopoulou *et al.*, 2017; Skehan, 2009; Tavakoli and Foster, 2011; Tavakoli and Skehan, 2005) as well as native speakers (e.g., Foster and Tavakoli, 2009; Pallotti, 2019; Staples *et al.*, 2016). Various definitions have been proposed for tasks and registers in research on (L2) writing, task-based language teaching and assessment, and SLA. In this thesis, I follow the broad notion of register proposed by Biber and Conrad (2001, p. 175, emphasis theirs): "Varieties defined in terms of general situational parameters are known as *registers*. We use the label *register* as a cover term for any variety associated with a particular configuration of situational characteristics and purposes". Further, I use Samuda and Bygate's (2008) task definition which encompasses all holistic activities that foster language learning by requiring learners to use linguistic means to reach a functional goal. Under this terminology, tasks are the controlled elicitation contexts that define the situational parameters which determine the (set of) appropriate register(s). Using this or similar definitions, the influence of several functional and cognitive task factors on language production has been investigated asking which task factors influence language performance and

how they interact with CAF and proficiency. Commonly investigated factors include genre and topic, cognitive demands (e.g., planning and reasoning, prior knowledge and familiarity, structure), social context and conditions (e.g., interactivity, formality, production mode, stakes, control), as well as linguistic demands (also known as 'code complexity', Skehan, 1998).

Several cognitive task-based frameworks have been proposed to jointly explain the influence of cognitive, functional, and situational task factors on L2 (and L1) performance and trade-offs within the CAF triad. Two of the most influential frameworks are Skehan's Limited Attention Capacity hypothesis (Skehan, 1998) and the Cognition Hypothesis by Robinson (1995, 2001). Both hypotheses assume that task demands and CAF tap into learners' cognitive resources and that these resources are limited. The competition between the CAF (sub-)dimensions is assumed to cause trade-off effects. The hypotheses make different predictions regarding these trade-offs primarily because Robinson's Cognition Hypothesis assumes that task demands may not only deplete cognitive resources but also re-direct them to amplify the resources allocated to specific CAF (sub-)dimensions. Robinson (2015) and Skehan (2015) provide a more detailed comparison of both hypotheses in a point-counterpoint discussion. Studies seeking to identify which of the two frameworks more accurately describes the effect of tasks on CAF yielded mixed results (Jackson and Suethanapornkul, 2013; Johnson, 2017).

Task effects have also been discussed from a functional perspective. There has been some work highlighting the influence of source material (Guo *et al.*, 2013; Kyle and Crossley, 2016; Miller *et al.*, 2016; Plakans and Gebril, 2013), discipline (Crossley *et al.*, 2017; Gardner *et al.*, 2019), and topic (Yang *et al.*, 2015; Yoon, 2017). However, genre effects have received special attention. Work on effects of genre (or 'task type', e.g. Alexopoulou *et al.*, 2017) has focused on the comparison of narratives, argumentative (or persuasive) texts (commonly essays), and expository (or informative) texts, but see Staples *et al.* (2016) for the comparison of more genre types in the advanced academic writing of native speakers at university. Few studies distinguish further between 'genre families' (e.g., letters, essays, reports; terminology from Staples *et al.*, 2016) and 'discourse mode' (e.g., narrative, expository, argumentative; terminology from Yang *et al.*, 2015). As Skehan (2009) points out, task factors have been mostly analyzed in terms of their influence on syntax. Work on genre effects for other linguistic domains is rather rare. Skehan (2009) summarizes six studies by Skehan and Foster that analyzed the effect of several cognitive task factors and genre (personal information exchange, narratives, decision making) on lexical complexity and sophistication. Also Yoon and Polio

(2017) and Alexopoulou *et al.* (2017) investigated genre effects on syntactic complexity, lexical complexity and accuracy (also fluency in the case of Yoon and Polio, 2017). Yoon and Polio (2017) found strong genre effects for lexical complexity in the sense that argumentative essays contained more sophisticated but less diverse vocabulary than narratives in both L2 and L1 writing. Olinghouse and Wilson (2013) studied the link between writing quality, genre, and lexical complexity in fifth graders English L1 writing, finding that the different genres fostered higher complexity in different sub-dimensions of lexical complexity. Tracy-Ventura and Myles (2015) found clear genre effects on morphology (specifically the use of past tense) in spoken L1 and L2 Spanish, contrasting the genres interview, narration, and description.

Despite these notable contributions, by far most work has been dedicated to the analysis of genre effects on syntactic complexity. One general and relatively stable finding across studies is that narrative writing elicits syntactically less complex writing than non-narrative writing (see also Lu, 2011; Yoon, 2017, p. 133). More specifically, non-narrative writing seems to promote phrasal complexity over clausal complexity: Beers and Nagy (2009) compared the link between essay quality ratings and clausal complexity for stories and argumentative essays written by English middle school pupils in seventh and eighth grade. The most notable difference that they found is that narratives elicited overall less elaborate clausal structures than argumentative essays. At the same time, clausal elaboration was positively correlated with quality ratings for narratives but negatively for quality ratings of argumentative essays. This ties into their second finding that phrasal rather than clausal elaboration is linked to higher quality in argumentative essays. Similarly, in their study on cross-disciplinary academic language development of English native speakers at university, Staples *et al.* (2016) found an increase of phrasal complexity and a decrease of clausal complexity from first-year undergraduate to graduate students (despite some discipline-specific differences). The same pattern seems to hold for L2 writing. Yoon and Polio (2017) found that English L2 learners' writing developed more complex phrasal structures in the course of four months in an instructed university setting but not more complex clausal structures. This effect was not present in L1 writing. Alexopoulou *et al.* (2017) analyzed English L2 texts elicited at different proficiency levels, observing that narrative texts by more proficient learners included more subordination and were overall more cohesive, whereas professional texts by more proficient learners contained longer clauses and more complex phrases.

The differences between narrative and non-narrative writing has some parallels to the observation that academic writing produces higher phrasal complexity (Bulté and Housen, 2014;

Lu, 2011; Yoon, 2017), whereas speech elicits more complex clausal structures (e.g., Biber *et al.*, 2011, 2016; Kormos and Trebits, 2012; Kuiken and Vedder, 2012). In fact, production mode has been shown to heavily influence language performance (Biber *et al.*, 2016; Kormos and Trebits, 2012; Kuiken and Vedder, 2012) and some evidence suggests that even the effect of task factors on complexity may vary across production modes (Kuiken and Vedder, 2012; Vasylets *et al.*, 2017). Biber *et al.* (2011) reason that this might suggest a developmental sequence from more complex clausal structures to more complex phrasal structures in native speakers' acquisition of academic writing skills because they acquire conversational skills prior to developing an advanced academic register. This idea is similar to the distinction of Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP) that was proposed by Cummins (1997) to distinguish two separate dimensions of L2 proficiency: conversational fluency and academic language use (for details, see Section 2.2.1, p. 44).

The findings from the studies discussed in the previous paragraph are in line with this hypothesis. The non-narrative writings were predominantly elicited by tasks defining situational parameters that elicit an academic language register. In contrast, the genre of narrative writing often utilizes characteristics associated with spoken language for stylistic reasons, especially in folk poetry (e.g., Dehrmann, 2014; Gerndt, 1988; Gobrecht, 1997; Seidenspinner, 1997), but also in other forms of prose (Mecklenburg, 2018; Tannen, 1982). Hence, it seems plausible that increasingly proficient narrative writing would employ more linguistic means associated with spoken language. Against this background, the frequently studied comparison of narrative and non-narrative genres seems to coincide with register differences (BICS/CALP). Unfortunately, the existing research on genre contrasts within non-narrative writing is insufficient to disentangle this confound.

Related to these functional approaches to task effects and concerns of register variation, the notion of functional adequacy (also sometimes referred to as communicative adequacy, e.g., Kuiken *et al.*, 2010; Pallotti, 2009) has received increasing interest in task-based language teaching and learning research (e.g., De Jong *et al.*, 2012; Hulstijn *et al.*, 2012; Kuiken *et al.*, 2010; Ortega, 2003; Pallotti, 2009). 'Adequacy' entails pragmatic and socio-linguistic concerns of appropriateness, effectiveness, and efficiency and task-based notions of communicative success and task completion. Pallotti (2009) argues that introducing adequacy into the notion of language performance allows to better interpret CAF measures. Specifically, it helps to account for the fact that "more complex" is not preferable in all communicative contexts as

discussed in Section 2.1.1.2 (p. 14). Highly proficient speakers of a language do not choose to maximize CAF (sub-)dimensions in all communicative contexts but to align their language with their communicative goals. Pallotti (2009) criticizes that SLA complexity research often attributes a lack of growth in one of the three dimensions (or their sub-dimensions, e.g., lexical versus syntactic complexity or clausal versus phrasal complexity) to trade-off effects. Instead, a lack of growth or even reduction along one (sub-)dimension might simply be functionally more adequate.[2]

Different approaches have been proposed to determine the functional adequacy of language productions. Pallotti (2009) proposed to use native speakers' language performance as benchmark. Similar proposals have been made to better differentiate between proficiency and task effects (Foster and Tavakoli, 2009; Skehan, 2009) and to quantify the linguistic demands of tasks (Pallotti, 2019). However, focusing on adequacy in terms of native speaker performance faces two challenges. First, native speakers can exhibit substantial inter-individual variation making it difficult to derive a stable baseline (Shadrova *et al.*, 2021). Second, even when a stable baseline can be obtained, it fails to identify potentially successful alternative strategies that learners might invent to compensate for a lack of linguistic means that native speakers do not need to face (see discussion in Pallotti, 2019). Human ratings have been proposed as an alternative to empirical benchmarks (Kuiken and Vedder, 2022; Pallotti, 2009; Révész *et al.*, 2014). Recently, Kuiken and Vedder (2017, 2018) introduced a multi-faceted holistic rating scale to measure adequacy based on task requirements, content, comprehensibility, as well as coherence and cohesion. It has been successfully applied (with some modifications) in several empirical studies and across languages (e.g., De Meo *et al.*, 2019; Pallotti, 2017; Révész *et al.*, 2014), see Kuiken and Vedder (2022) for an overview. Yet, there are so far no large-scale corpora containing functional adequacy ratings that could be used for computational linguistic research.

2.1.3.3 Cross-lingual variation

Linguistic typology has studied synchronic and diachronic variation in language systems along the dimension of complexity for decades (for an overview, see Dahl, 2004; Karlsson *et al.*, 2008). Beyond comparing different languages to each other in terms of their absolute or relative complexity, language change and the role of language contact is a dominant theme in this line of research (see Miestamo *et al.*, 2008, and contributions therein). There have been some

[2] Similar arguments can be made for accuracy and fluency, see Pallotti (2009, p. 597).

attempts to directly compare the global complexity of languages, for example with measures from information theory (Dahl, 2004; Ehret, 2018; Ehret and Szmrecsanyi, 2016). However, quantifying the global complexity of a language in a way that lends itself to a linguistically meaningful interpretation remains an open challenge (see Section 2.1.1.2). Thus, most cross-linguistic analyses have focused on local complexity and the comparison of specific functional domains of grammar that are comparable across languages (Miestamo, 2008, p. 23).

For decades, languages were hypothesized to be equally complex on a global level. The so called 'equi-complexity hypothesis' postulates that reduced complexity in a specific linguistic domain (e.g., morphology) is compensated through higher complexity in another linguistic domain (e.g., lexicon), rendering all languages equally complex. All languages were argued to have the same functional and communicative needs. These needs should determine an appropriate degree of complexity that no language should fall below or exceed. The equi-complexity hypothesis has been heavily challenged on methodological and empirical grounds since the early 2000s (e.g., Karlsson *et al.*, 2008; Kusters, 2008; McWhorter, 2001; Miestamo, 2008). Today, researchers predominantly agree that languages vary considerably in terms of their complexity, not only in their different linguistic domains but also globally (e.g., Karlsson *et al.*, 2008; Kusters, 2008; Miestamo, 2008). Beyond language community size (Lupyan and Dale, 2010; Reali *et al.*, 2014), the degree of language contact and L2 acquisition that a language is being exposed to have been identified as two important sources of language change and complexity differences between languages. Specifically, several studies identified high degrees of language contact and the influence of L2 learners as simplifying influences (Bentz and Berdicevskis, 2016; Karlsson *et al.*, 2008; Lupyan and Dale, 2010; McWhorter, 2008), but see De Groot (2008) for an opposing view. Languages that are isolated from simplifying influences have been argued to become more complex over time (Dahl, 2004; McWhorter, 2001). This is in line with Rescher's (1998) assumption that the complexity of systems grows over time.

Cross-linguistic variation is also a relevant concern for complexity research focused on quantifying learners' language development and proficiency. Most research on CAF has focused on English assuming that findings would transfer to other languages (De Clercq and Housen, 2017, p. 319). However, there has been some evidence of cross-linguistic differences in the development of learner language (e.g., Brezina and Pallotti, 2019; Kuiken and Vedder, 2012; Martin *et al.*, 2010). To address the question how first and second language development in other languages may differ from previous findings for English, two main approaches

have been explored. There have been several studies focusing on the development of linguistic complexity for languages other than English (e.g., Kuiken and Vedder, 2012; Martin *et al.*, 2010; van der Slik *et al.*, 2019; Vyatkina *et al.*, 2015). However, the comparison of studies focusing on different languages is often challenging. In addition to the common obstacles to study comparisons—such as the use of different measures and differences in task setting or study populations—comparisons across languages face the additional question of the extent to which features have been operationalized in a comparable manner or are conceptually comparable. This concern has been addressed through partial replication studies of findings reported for English (e.g., Rubin, 2021; Vandeweerd *et al.*, 2021) or through multi-lingual research designs. For example, a series of the cross-lingual developmental variation of morphological complexity found that the inflectional diversity of English L1 and L2 samples was systematically lower than the inflectional diversity of Italian (Brezina and Pallotti, 2019) or French (De Clercq and Housen, 2019) L1 and L2 samples. There is also evidence that the developmental variation of the morphological diversity in L2 samples is less pronounced and levels off at earlier developmental stages for English than for morphologically richer languages (Brezina and Pallotti, 2019; De Clercq and Housen, 2019)

Cross-lingual variation is also an important factor to consider in the context of task effects, register variation, and functional adequacy (see Section 2.1.3.2, p. 37). Kuiken and Vedder (2012) found that cognitive task effects on lexical sophistication differed in French and Italian L2 speech. Functional task factors, too, can be subject to cross-linguistic differences. As De Clercq and Housen (2017, p. 319) argued, language communities may differ in their stylistic preferences and rhetorical strategies (see also Fausey and Boroditsky, 2011). A well-studied example for such differences are different preferences across academic communities (see e.g., Pallotti, 2009, p. 598). For example, the previously discussed shift towards high phrasal complexity and a reduction of clausal complexity for English academic language (Biber *et al.*, 2011, see also Section 2.1.3.2, p. 36) does not generalize to German academic language. German *Bildungssprache* (engl. "academic language") has been characterized on the syntactic level as highly elaborate in terms of its clausal structures as well as its phrasal complexity (Hawlik and Sorger, 2017; Stahns, 2016). These cross-lingual differences not only play a role for SLA complexity research because they question to what degree findings for English transfer to other languages. They also matter in terms of L1 transfer effects on L2 productions (De Clercq and Housen, 2017). Against this background, it becomes clear that more research on cross-lingual variation and its role for language learning is needed. It

is important to promote complexity research on languages other than English. Research also needs to facilitate the broad linguistic comparison of languages in terms of local complexity measures that have been operationalized in a comparable way. This is needed to foster generalizable insights into the link between complexity and proficiency and to inform language teaching and learning practice.

2.2 Automatic proficiency assessment

Automatic proficiency assessment seeks to approximate a learners' proficiency by analyzing a freely produced language sample (Vajjala, 2018; Yannakoudakis and Cummins, 2015). The assessment can be holistic or focus on specific dimensions of proficiency (such as vocabulary knowledge, grammatical control, or functionally adequate language use). To quantify the proficiency estimate, automatic proficiency assessment utilizes a predefined rating scale. This scale may quantify a large range of proficiency (e.g., beginning to advanced L2 proficiency) or focus on performance nuances in a more narrowly defined proficiency range (e.g., grades, fail/pass). This section provides a focused background on automatic approaches to proficiency assessment in computational linguistics. It thus complements Section 2.1.3.1, which discusses SLA research on developmental variation of complexity. Computational linguistic work on automatic proficiency assessment typically assesses prompt-based writing (Hussein *et al.*, 2019; Vajjala, 2018), even though there has been some limited work on the assessment of speech data (Bhat and Yoon, 2015; Xie *et al.*, 2012; Zechner *et al.*, 2009). In view of this strong focus on written language, the terms Automatic Text Scoring (ATS) or—even more narrowly—Automatic Essay Scoring (AES) are frequently used in computational linguistics. In the following, I will use the term ATS instead of automatic proficiency assessment to follow these terminological conventions. The remainder of this section is structured as follows. I will briefly elaborate on the term 'proficiency' as it is used in the context of language learning, teaching, and assessment (Section 2.2.1). I will then describe the central application domains for ATS (Section 2.2.2) before focusing on the current methods and trends in ATS research (Section 2.2.3).

2.2.1 What is language proficiency?

In SLA complexity research, the notion of 'language proficiency' is typically not—or only loosely—defined despite its relevance for developmental variation and its use as proxy for

language development in cross-sectional studies (see Section 2.1.3.1). Generally, it is used to quantify the current state of someone's acquisition process in relation to a functional goal or a reference standard (e.g., Peña *et al.*, 2021, p. 89). It is thus a latent variable that cannot be directly observed but than can be approximated through language performance. It differs from the concept of 'language development' which views the process of language acquisition from a longitudinal perspective. In the few cases where proficiency is explicitly defined in SLA complexity studies, it is viewed from a general perspective without elaborating on its (sub-)dimensions or referencing existing theoretical frameworks. For example, proficiency has been referred to as "overall competence and ability to perform in L2" (Thomas, 1994, p. 330, Footnote 1), see also Bulté and Housen (2014).

In contrast, in language testing research, several theoretical models have been proposed to define language proficiency and its (sub-)dimensions. One influential contribution is the theoretical framework of 'communicative competence' proposed by Canale and Swain (1980) and extended by Canale (1983). It considers not only learners' command of vocabulary and grammar but also their sociolinguistic, strategic, and discourse competence. Extending on this notion of communicative competence, Bachman and Palmer (1996) proposed a model of language proficiency that focused on test performance. It considers not only learners' grammatical, pragmatic, and strategic competence but further incorporates the influence of learners' individual characteristics (e.g., topic knowledge, age, gender, L1s, L2s). Many more theoretical frameworks have been proposed in language testing research (for overviews, see, e.g., Piggin, 2012). The various theoretical models of language performance differ in how they weigh the role of linguistic knowledge, processing ease, and pragmatic or meta-linguistic knowledge for language proficiency. Hulstijn (2011) argues that linguistic knowledge and processing ease should be considered at the core of language proficiency. Meta-linguistic and strategic competencies are peripheral dimensions of language proficiency because they require linguistic knowledge but not vice versa (Hulstijn, 2011, pp. 238–239). This is in line with the focus of SLA research on CAF as dimensions of language performance while accounting for the dimension of functional adequacy discussed previously (Section 2.1.3.2, p. 2.1.3.2).

Within educational contexts, language proficiency is commonly expressed in terms of levels. This provides a convenient way to quantify proficiency for language teaching and assessment. In language teaching, it allows to group learners of similar proficiency together and provide them with a matching curriculum. Similarly, in language assessment, the notion of proficiency levels allows to match test takers with tests of appropriate difficulty and

avoid ceiling or floor effects that would invalidate the assessment. However, proficiency levels are primarily practical categories. They are usually not intended as conceptually grounded components of a theoretical proficiency framework. "Any attempt to establish 'levels' of proficiency is to some extent arbitrary [...] However, for practical purposes it is useful to set up a scale of defined levels to segment the learning process for the purposes of curriculum design, qualifying examinations, etc" (Council of Europe, 2001, p. 17). With the Common European Framework of Reference for Languages (CEFR), the Council of Europe (2001) has defined one of the most important frameworks for the definition of language proficiency levels. It defines a vertical scale of six consecutive proficiency levels which are assumed to align with the developmental trajectory of learners (Council of Europe, 2001, p. 17). Basic learners are distinguished into the levels A1 (breakthrough), A2 (waystage), independent users into the levels B1 (threshold) and B2 (vantage), and proficient users into the levels C1 (effective operational proficiency) and C2 (mastery), see Council of Europe (2001, pp. 33–36). Note that these levels are intended for L2 learners. The level C2 is "not intended to imply native-speaker or near native-speaker competence" (Council of Europe, 2001, p. 36).

Each CEFR level is further specified along a horizontal dimension. This dimension considers communicative language competence (divided into linguistic, sociolinguistic, and pragmatic components as in Canale and Swain, 1980) and language activities (including reception, production, interaction, and mediation). Language activities are designed to allow users to perform their communicative competencies across usage contexts (public domain, personal domain, educational domain, occupational domain) by using different strategic competencies (Council of Europe, 2001, pp. 13–16). Language development is assumed to occur both in the vertical and the horizontal dimension of the CEFR scale (Council of Europe, 2001, p. 17). This implies that learners at the same CEFR level can exhibit different proficiency profiles in terms of their breadth and width along the vertical and horizontal dimension. Some researchers have criticized that the CEFR levels should be more clearly separated from the notion of L2 development because they do not reflect empirically grounded developmental sequences (e.g., Alderson, 2007; Deygers, 2021; Hulstijn, 2007; Wisniewski, 2017). Also Hulstijn (2011) criticizes the confound of proficiency and development in the CEFR scale. He argues that the CEFR levels do not define developmental trajectories because the B2 to C2 levels are not attainable for all L2 learners: "A close examination of the definitions of the B2, C1, and C2 levels in the activity and the competence scales reveals that performance at these higher levels requires higher *intellectual skills*" (emphasis his, Hulstijn, 2011, p. 241) and are associated

with higher education. Hulstijn further argues that "many adult native speakers will never attain the highest CEFR levels (C1 and C2)" for the same reasons (Hulstijn, 2011, p. 241, see also pp. 240–241 for full discussion).

Multiple frameworks advocate for the distinction of two types of language proficiency: one linked to basic or conversational language use and one (typically written) linked to sophisticated or academic language use (e.g., Gee, 1990; Gibbons, 1991). Because the former is typically associated with spoken language and the latter with predominantly with written language, Cummins (2000) draws also parallels between these frameworks and work on register differences between spoken and (academic) written language (e.g. by Biber, 1986). One of the most prominent proposals making this distinction is Cummins' (1997; 2008) who distinguishes between Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP). BICS refer to the language skills that are acquired through social interactions outside of academic contexts. BICS are predominantly oral and native speakers reach high proficiency in this dimension at an early age. In contrast, CALP refers to written and spoken academic language proficiency. While L2 or L1 speakers may develop linguistic knowledge associated with CALP prior to entering school, Cummins assumes that CALP begins to manifest as a separate dimension of language proficiency in the first years of schooling and continues to develop through exposure to academic contexts (Cummins, 2000, 2008). The distinction of BICS and CALP was originally made to explain seemingly contradictory observations in early L2 acquisition and bilingual pupils (for overview, see Cummins, 2008) but has also been used in language teaching research to discuss native speakers' early academic language acquisition in school (e.g., Weiss *et al.*, 2021).

Hulstijn (2011, 2015) proposed a similar proficiency distinction to provide a unified framework of language proficiency that can be applied to L2 contexts while also accounting for individual differences in native speakers' language command (Hulstijn, 2011, p. 229). Hulstijn defines language proficiency as "the extent to which an individual possesses the linguistic cognition necessary to function in a given communicative situation, in a given modality (listening, speaking, reading, or writing)" (Hulstijn, 2011, p. 242). Before discussing the two forms of proficiency distinguished in this framework, I briefly contextualize the notion of 'linguistic cognition'. It is a two-dimensional construct, consisting of a) explicit and implicit phonological, morphological, lexical, syntactic, semantic, and pragmatic knowledge as well as b) the ease of processing linguistic information (Hulstijn, 2011, p. 242). Linguistic knowledge is at the heart of this notion of linguistic cognition but it also encompasses strategic and

meta-cognitive competencies as a peripheral component (Hulstijn, 2011, p. 242). Hulstijn distinguishes between two types of linguistic cognition: Basic Language Cognition (BLC) and Higher Language Cognition (HLC). BLC refers to the linguistic knowledge and processing skills shared between native speakers independent of their education, literacy, or age (except the first years of life). It focuses on spoken language, everyday communicative needs, and frequent vocabulary and morpho-syntactic constructs. It is similar to Cummins's notion of BICS but Hulstijn points our that his notion of BLC is more fine-grained because Cummins's framework focuses on CALP (Hulstijn, 2011, pp. 232–233). HLC (similar to Cummins's CALP) refers to the additional knowledge and processing skills needed to use and comprehend less frequent vocabulary and morpho-syntactic constructs. HLC is used for academic language in spoken or written form in private, academic, or professional settings. Hulstijn (2011, 2015) assumes that individual differences in native speakers' language proficiency are more prevalent in HLC than in BLC (Hulstijn, 2011, pp. 230–232). Differences in L1 and L2 performance are assumed to be most pronounced in terms of BLC. In fact, Hulstijn (2011, p. 242) states that "L2 learners can be as proficient in HLC as L1-ers of the same intellectual, educational, professional, and cultural profile, despite some deficiencies in their L2 BLC." Hulstijn's and Cummins's distinctions between basic, communicative language proficiency and academic language proficiency are particularly useful for this thesis, for example in Weiss and Meurers (2019a, Section 5.2.3). They facilitate not only the discussion of advanced academic language development in native speakers and individual differences in their reception of language. They also account for some of the register and task factors on language performance along the dimension of complexity that were discussed in Section 2.1.3.2.

2.2.2 Application domains

ATS is one of the most common applications of computational linguistics in education (Ke and Ng, 2019). In this section, I distinguish the three main application domains of ATS in education contexts. These are i) summative assessment, ii) formative feedback, and iii) research on writing quality. For a systematic review focused on ATS for German, see Section 4.2.

2.2.2.1 Summative assessment

The most straightforward and common application domain of ATS is the assessment of language proficiency or writing quality. This application focuses on evaluating the writing prod-

uct instead of the writing process (for a discussion, see Quinlan *et al.*, 2009, pp. 3–5). ATS seeks to approximate human ratings of task performance or proficiency in open answer formats. These are commonly essays, hence the commonly used, more narrow term Automatic Essay Scoring (AES). However, also short answers can be assessed using ATS. Here, the focus typically lies on the assessment of the answer contents rather than the language use (Burrows *et al.*, 2015). Short answer assessment and AES are therefore using very different methods despite both being types of ATS. Open answer formats—also known as constructed response items (Chen *et al.*, 2016)—are a central components in many high-stakes standardized language proficiency tests such as the GRE or TOEFL (Crossley *et al.*, 2016b; Vajjala, 2018). However, human expert ratings are time-consuming and costly. They are also prone to several biases (e.g., severity/leniency, scale shrinkage, inconsistency, halo effects, stereotyping, perception difference, and rater drift; cf. Zhang, 2013, p. 2), see also, e.g., Kassim (2011); Myford and Wolfe (2003). To alleviate these issues, several quality standards have been introduced, including standardized scoring methods and double-scoring (McClellan, 2010; Quinlan *et al.*, 2009). This in turn has further raised costs associated with human ratings.

This has made ATS a crucial tool in language instruction and large-scale language assessments and several professional ATS systems have been on the market since the 1990s. For a list of early systems, see Attali and Burstein (2006, p. 3) or Chodorow and Burstein (2004, p. 1). For a more recent overview of ATS systems, see Hussein *et al.* (2019). In practice, ATS systems are not recommended to be used in high-stakes testing without human supervision (Powers *et al.*, 2002; Zhang, 2013) because of their sensitivity to adversarial input (Powers *et al.*, 2002) and their insensitivity to important aspects of writing such as originality and style (Attali, 2007, p. 1). However, they can substitute human raters in large-scale, low-stakes contexts (Chodorow and Burstein, 2004, p. 6; Zhang, 2013, p. 6), for example for writing training or to match learners with competence-adaptive learning materials (Chen and Meurers, 2019). They have also been used for decades in high-stakes contexts in two ways. First, to reduce the number of human expert raters needed for double-scoring in language assessments by replacing one human rater (Powers *et al.*, 2002; Quinlan *et al.*, 2009; Vajjala, 2018; Zhang, 2013). This is known as ‘contributory scoring’ (Chen *et al.*, 2016, p. 3). Second, to control the quality of human ratings (Monaghan and Bridgeman, 2005; Wang and von Davier, 2014; Zhang, 2013). This can for example be achieved by using ATS systems to flag essays for which human ratings and automatic ratings deviate above a predefined threshold, so that they can be inspected by an additional human rater. This is known as ‘confirmatory scoring’ (Chen *et al.*,

2016, p. 3).

One of the most prominent and established ATS systems is the e-rater Scoring Engine (short: e-rater; Attali and Burstein, 2006, `https://www.ets.org/erater`). e-rater is a proprietary AES system developed by the Educational Testing Service (ETS). It has been used for large-scale assessments since 1999 (Chen *et al.*, 2016, p. 2) in the Graduate Management Admission Test (Attali and Burstein, 2006, p. 4; Chodorow and Burstein, 2004, p. 6). Since then, it has been used repeatedly for contributory and confirmatory scoring in high-stakes testing (Chen *et al.*, 2016; Ramineni and Williamson, 2018) and as stand-alone rater in low-stakes contexts in schools (Chodorow and Burstein, 2004, p. 6). e-rater is a feature-based machine learning system (Attali and Burstein, 2006; Chen *et al.*, 2016) that focuses on essay scoring. It provides holistic scores as summative feedback, but has also a formative feedback component (Quinlan *et al.*, 2009, p. 5). e-rater has been designed to make predictions based on eight to 12 (macro-)features (Chen *et al.*, 2016, p. 3; Powers *et al.*, 2002, p. 5) that cover meaningful components of human writing quality estimates while avoiding pure surface-based measures of text length (Burstein and Chodorow, 1999, p. 69). These features are partially assessed through aggregating several so called sub-features and micro-features. They focus on measuring essays' structure, organization, and content (Powers *et al.*, 2002, p. 5) through a combination of syntactic, lexical, and discourse complexity features, as well as accuracy measures and prompt-specific vocabulary and topic features (Attali and Burstein, 2006, p. 11; Chen *et al.*, 2016, p. 3; Powers *et al.*, 2002, p. 5). The system relies on a small set of aggregate features that can be meaningfully related to important dimensions of writing quality to promote the face validity and interpretability of scores (Attali and Burstein, 2006; Quinlan *et al.*, 2009) as well as to support formative feedback along meaningful dimensions (Attali and Burstein, 2006).

Additionally, e-rater utilizes a separate analysis component that seeks to identify adversarial essays (Attali and Burstein, 2006, p. 12; Powers *et al.*, 2002, p. 18) based on measures of essay-prompt overlap (to find off-topic essays) and intra-text repetition (to find paragraph duplication). When applied to a new assessment context, e-rater can be used as a generic cross-prompt system or be adapted to the new application domain (Attali and Burstein, 2006; Quinlan *et al.*, 2009) either through data-driven or theory-driven methods (Attali and Burstein, 2006; Powers *et al.*, 2002). Data-driven domain adaptation relies on learning feature weights from domain-specific training data. Theory-driven domain adaptation relies on human expert judgments regarding the relevance of the individual features for the given rating context. The system also supports hybrid approaches, for example by constraining data-driven feature

weights based on theoretical considerations (Attali and Burstein, 2006). In training, e-rater is designed to prioritize precision over recall (Quinlan *et al.*, 2009, p. 14). e-rater has been shown to provide robust and reproducible results, with adjacent accuracies ranging between $90-95\%$ (Chodorow and Burstein, 2004, p. 6; Powers *et al.*, 2002, p. 9). It has also been evaluated in several validation studies (Attali, 2007; Attali and Burstein, 2006; Chodorow and Burstein, 2004; Powers *et al.*, 2002; Quinlan *et al.*, 2009), testing for example its robustness against adversarial input (Powers *et al.*, 2002) and its independence from surface text characteristics such as essay length (Chodorow and Burstein, 2004).

Open answer formats are also recognized in other educational contexts as valuable instruments for assessing higher levels of understanding and advanced competencies that cannot be readily measured with closed answer formats (e.g., fill-in-the-blank, yes/no questions, multiple choice questions), see Esses and Maio (2002); Schuwirth and Van Der Vleuten (2004); Smith *et al.* (2019), especially in humanities. Also in these application contexts, ATS is being promoted as a resource-efficient and robust alternative for assessment contexts in which closed answer formats are unsuited and human expert ratings not feasible (Hussein *et al.*, 2019; Uto, 2021; Vajjala, 2018). However, outside of language assessment and testing contexts, content-based ATS has been much more prominent, for example in form of short answer assessment (for an overview, see Burrows *et al.*, 2015; Ziai, 2018). Content-based ATS is less focused on aspects of language use and performance and often use reference answers as gold standard (Padó, 2016; Vajjala, 2018). For some tasks or advanced proficiency levels, the functional adequacy of language use has been considered as an additional performance dimension outside of language learning, too (e.g., Frey, 2020a; Ludwig *et al.*, 2021). Bertram *et al.* (2021) recently explored the feasibility of a hybrid system that assesses language productions in terms of their content and their functionally adequate language use for history education.

To a lesser degree, ATS systems for the assessment of writing quality have also been proposed outside of education. Examples are content quality assessment for collaborative writing platforms (mostly Wikipedia, see Hasan Dalip *et al.*, 2009; Shen *et al.*, 2017, 2019) or the automatic review of scientific papers (Deng *et al.*, 2020; Leng *et al.*, 2019; Lin *et al.*, 2021). These approaches are methodologically very close to summative ATS. However, they can also include aspects of ARA (Section 2.3) depending on the degree to which readability is a component of writing quality for the specific evaluation context. There has been relatively little work exploring this connection between text quality and readability, but see Chen and Meurers (2019) for an approach focused on language learning. Others have investigated links between

readability and scientific or economic success (Ashok *et al.*, 2013; McCannon, 2019; Shelley and Schuh, 2001).

2.2.2.2 Formative feedback

While the previously discussed use of ATS systems focused on language assessment, ATS can also be used in language learning and teaching. Unlike human raters, ATS systems can provide learners with direct formative feedback while writing to support their writing process (Zhang, 2013), see Hussein *et al.* (2019) for a survey of available systems. These types of ATS systems are sometimes referred to as Automatic Writing Evaluation (AWE) systems to better distinguish them from systems focusing on summative feedback and assessment (Crossley, 2020, p. 416). AWE systems have also been embedded into intelligent tutoring systems, for example Writing-Pal (Roscoe *et al.*, 2014) which additional to writing practice with feedback also provides targeted instructions on writing strategies. Formative writing feedback in AWE systems focuses on writing as a process. It seeks to foster the accuracy, cohesion, and overall discourse structure of writing, as well as promoting its functional adequacy in form of task-orientation and register awareness (Crossley, 2020; Hussein *et al.*, 2019). To do so, AWE systems provide ratings on sub-dimensions of writing quality and proficiency instead of or additional to holistic ratings (Quinlan *et al.*, 2009, p. 5). Several researchers have pointed out the relevance of AWE for online learning and teaching—especially for Massive Open Online Courses (MOOCs)— for learners (Vajjala, 2018, p. 80) and educational data mining (Crossley *et al.*, 2015, p. 204).

The eRevis system (Zhang *et al.*, 2019) is an example for such an AWE. It focuses on promoting the use of text evidence on argumentative source-based writing for young learners (Wang *et al.*, 2020). It assesses evidence use in terms of four features which were designed to approximate the 'text evidence' rating rubric for Response-to-Text Assessment (Wang *et al.*, 2020, p. 4). The features consists of the total `word count` as well as three features that compare students' vocabulary with the source material. Relevant vocabulary is identified using skip-grams and topic modeling on the source data (see Rahimi and Litman, 2016; Zhang *et al.*, 2019). These features aim to quantify how much text-based evidence students used as well as the vocabulary specificity and evidence density of the text (for details, see Rahimi and Litman, 2016; Rahimi *et al.*, 2014; Zhang *et al.*, 2019). eRevis has been shown to perform close to human raters in terms of inter-rater reliability (IRR) on a corpus of source-based writing essays of 5th and 6th grade students (Correnti *et al.*, 2020). Correnti *et al.* (2020)

further externally validated the model's predictions by comparing the consistency between the automatic rating and other performance scores of students. Based on its assessment, eRevis provides students with four formative feedback levels, prompting students to provide more evidence, details, explanations, or connections. These four messages are paired into three feedback levels. Low scoring essays are prompted to focus their revision on incorporating more evidence and details. Medium scoring essays are prompted to provide more details and explanations. High scoring essays are advised to focus their revision on more elaborate explanations and connecting arguments further. The feedback levels are chosen based on a feature-driven feedback selection algorithm (for details, see Zhang *et al.*, 2019). Wang *et al.* (2020) piloted the effectiveness of the feedback for students in 5th and 6th grade. They report that students' argumentative essays mostly improved based on revisions prompted by the eRevis system even though the total improvement was small.

Stevenson and Phakiti (2014) conducted a survey on AWE systems and the effectiveness of their formative feedback. They found only modest evidence that AWE systems and their automatic feedback substantially improve writing products beyond reducing the number of errors. Stevenson and Phakiti (2014) reported several methodological weaknesses and too optimistic interpretations of mixed results. They advocate for more research on AWE systems and their effectiveness, focusing on more controlled experimental settings and the comparison of the impact of automatic feedback with teacher feedback. More research is also needed on how to obtain robust and valid quality estimates along specific (sub-)dimensions of proficiency (Ke and Ng, 2019; Uto, 2021).

2.2.2.3 Writing quality research

Work on ATS has also facilitated research on writing quality and language proficiency. This holds especially for feature-based approaches to ATS (see Section 2.2.3). Studies on how to approximate human judgments of writing quality and language proficiency have not only yielded insights into the link between language development, human proficiency or quality ratings, and linguistic text characteristics (Crossley, 2020, p. 416), which as been identified as an important research objective in SLA complexity research (Bulté and Housen, 2014, p. 43). They have also promoted the design of systems that automatically extract complexity and accuracy measures such as Coh-Metrix (Graesser *et al.*, 2004; McNamara *et al.*, 2010a), L2SCA (Lu, 2010), TAACO (Crossley *et al.*, 2016c), or CTAP (see Section 3.3 Chen and Meurers, 2016; Weiss *et al.*, 2021). More details on such systems can be found in this thesis

in Section 3.1.1. These—or similar tools—have in turn been used in research on writing quality (for an overview, see Crossley, 2020) and research on L2 development (e.g., Crossley *et al.*, 2010a; Crossley and McNamara, 2014; Crossley *et al.*, 2011a; Lu, 2011, 2012; Yoon, 2018; Yoon and Polio, 2017). For a more detailed discussion on this line of research, see Section 2.1.3.

2.2.3 Current methods and trends

Research on ATS dates back to the 1960s (Page, 1966, 1968; Whalen, 1971) with strong ties to writing quality and language learning research (Crossley, 2020). Advances in statistical NLP have in the early 2000s further promoted the development of increasingly sophisticated ATS systems. Most research on ATS has focused on English (Crossley, 2020; Vajjala, 2018). Despite research on ATS for other languages (e.g., Berggren *et al.*, 2019; Caines and Buttery, 2020; Hirao *et al.*, 2020; Östling *et al.*, 2013; Weiss and Meurers, 2019b), real life applications of ATS—especially in language testing—are near-exclusively available for English (Vajjala, 2018, p. 82). In the following, I will discuss the current methods and trends in supervised ATS in terms of machine learning tasks and evaluation metrics (Section 2.2.3.1) as well as rating scales and corpora (Section 2.2.3.2) used. I then compare neural and feature-based approaches and discuss the type of linguistic features used for ATS (Section 2.2.3.3).

2.2.3.1 Machine learning tasks and model evaluation

The majority of machine learning-based ATS approaches uses supervised methods (Ke and Ng, 2019) and relies on human judgments of proficiency or text quality as gold standard labels for training (Crossley *et al.*, 2016b; Vajjala, 2018). For two notable exceptions, see Chen *et al.* (2010) and De and Kopparapu (2011). Supervised ATS is typically framed as a classification or regression task (Borade and Netak, 2020; Hussein *et al.*, 2019; Ke and Ng, 2019) but some studies have also used (pair-wise) ranking (Yannakoudakis *et al.*, 2011). The use of classification algorithms in ATS requires a brief explanation, as they disregard the inherently ordered nature of rating scales. The choice of machine learning algorithm is closely linked to the properties of the proficiency scale on which a model is trained. Proficiency ratings are typically ordinal or discrete numerical data with a limited range (such as 1–6 or A–F). Even when data is represented in numerical form, they may be ordinal in the sense that adjacent rating categories are not guaranteed to be equidistant. Consider for example a 4-point Likert

scale distinguishing between 'very good' (1), 'good' (2), 'medium' (3), and 'insufficient' (4). The difference that human raters make between 1 ('very good') and 2 ('good') can be smaller than the difference between 3 ('medium') and 4 ('insufficient'). These properties of proficiency ratings make classification algorithms a popular alternative to regression algorithms. Similarly, pair-wise ranking approaches avoid the assumption of equidistance, but at the cost of not being able to assign an absolute performance score.

Depending on the choice of machine learning algorithm, different evaluation metrics are used. Regression models are typically evaluated in terms of correlation coefficients often in form of Pearson correlation or Spearman's rank correlation (Yannakoudakis and Cummins, 2015). However, these measures only indicate if the predicted scores and gold standard ratings are in a linear relationship; they do not quantify the actual agreement. They are thus 'measures of association' not 'measures of agreement' as Yannakoudakis and Cummins (2015) put it. Commonly used agreement estimates for regression-based systems are root mean squared error (RMSE) and mean absolute error (MAE). Classification approaches are typically evaluated in terms of accuracy (Vajjala, 2018) as well as precision, recall, and f-score (Borade and Netak, 2020; Vajjala, 2018). Many studies report not only the exact accuracy of their system but also its 'adjacent accuracy' to account for the fact that classification errors between adjacent levels are less severe than classification errors between more distant levels (see for example Crossley *et al.*, 2015, 2016b; Vajjala, 2018). A mathematically more formal way of accounting for the different weighting of classification errors is the use of weighted Cohen's kappa, typically in form of quadratic weighted Cohen's kappa. Cohen's kappa (Cohen, 1960) is a metric of chance-corrected IRR which is typically used to estimate the agreement between human annotators—or in the case of ATS systems between the automatic and the human rating. Its weighted version (Cohen, 1968) penalizes certain disagreements (for example between related or adjacent labels) less severely than others. Using measures of IRR to evaluate ATS also facilitates the comparison between human raters and ATS systems (as done for example in Wahlen *et al.*, 2020, p. 2). Human IRR is often used as an important benchmark in the evaluation of ATS systems (Attali and Burstein, 2006; Quinlan *et al.*, 2009). This acknowledges that human ratings of open answer formats—the systems' gold standard—are more variable than ratings of closed answer formats (as discussed in Section 2.2.2, see also for a more detailed discussion Quinlan *et al.*, 2009, pp. 14–15). For these reasons, Cohen's kappa is often used to evaluate ATS approaches (Borade and Netak, 2020) and has been recommended in

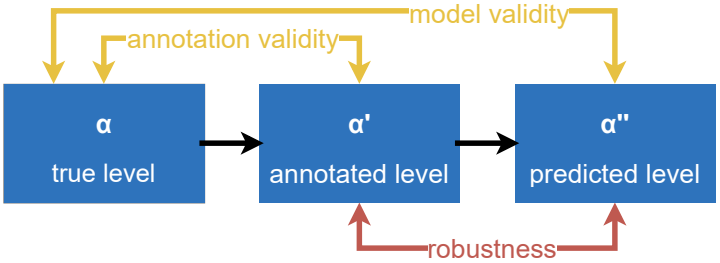


Figure 2.1: *Robustness, annotation validity, and model validity*

challenges such as the Automated Student Assessment Prize[3]. For an in-depth discussion of its mathematical properties in contrast to accuracy and correlation metrics, see Yannakoudakis and Cummins (2015).

The evaluation metrics discussed here focus on the 'robustness' of predictive models, that is how well they predict human scores used as training labels. Another important aspect to consider is 'construct validity', that is, how well the latent target construct (in this case proficiency level) is being approximated (Quinlan *et al.*, 2009, p. 2). This is a particularly important concern for educational and high-stakes contexts. For supervised machine learning, construct validity is essential for two estimates: the model's prediction (henceforth 'model validity') and the human annotation (henceforth 'annotation validity'), see Figure 2.1. Annotation validity refers to the agreement between the latent proficiency level α and the annotated reference proficiency level used for training a model α'. Model validity refers to the agreement between α and the proficiency level predicted by the ATS model α''. In supervised machine learning settings, model validity depends (among other factors) directly on the validity of the gold standard annotations used for training, making α'' dependent on α'. If α' provides an insufficient approximation of the underlying latent variable, the model cannot learn a valid characterization of the latent variable from the annotations. This makes it crucial to question the quality of training labels when assessing the quality of a model that aims to predict a latent variable. There is abundant work on inter-rater reliability and the construct validity of human ratings from language testing and writing quality research. This includes, for example, research investigating the validity of the CEFR scale (Hulstijn, 2014; Wisniewski, 2011, 2017), studies comparing language development and human proficiency ratings (Crossley and McNamara, 2014; Crossley *et al.*, 2016a), and experiments testing human raters' sensitivity to linguistic differences in essays (Vögelin *et al.*, 2019; Weiss *et al.*, 2019).

[3] `https://www.kaggle.com/competitions/asap-aes/overview/description`

There has also been considerable work on testing model validity of proprietary systems such as e-rater (see e.g., Attali, 2007; Attali and Burstein, 2006; Powers *et al.*, 2002; Quinlan *et al.*, 2009), but it is less commonly controlled for in research contributions (Ke and Ng, 2019; Uto, 2021). Quinlan *et al.* (2009) advocate for basing the ATS system on features that broadly measure construct relevant aspects of writing quality of proficiency (Quinlan *et al.*, 2009, p. 23), see also Attali and Burstein (2006) for a similar argument. This addresses two issues simultaneously: construct under-representation—not accounting for relevant aspects of writing quality may lead to the system underestimating the true essay quality—and construct irrelevance—relying on construct-irrelevant characteristics of writing such as text length may cause the system to overestimate essay quality—(Powers *et al.*, 2002, p. 3). A related validation strategy is to check model predictions for their robustness against adversarial input, such as off-topic or nonsensical essays as well as text repetitions (for a discussion of adversarial strategies and their impact on ATS systems, see Powers *et al.*, 2002). A model that suffers neither from construct-irrelevance nor construct under-representation should provide a valid representation of the underlying construct and therefore also yield robust performance on attempts to cheat the system (Powers *et al.*, 2002, p. 3). Finally, several researchers have proposed to focus more on extrinsic evaluation methods to ensure models' construct validity. Attali and Burstein (2006) proposed to compare the consistency between model predictions and other estimates of writers' proficiency, such as scores for other essays written by the same person for different prompts (Attali and Burstein, 2006, p. 5), see Correnti *et al.* (2020) for a similar argument. Ke and Ng (2019) proposed to conduct more user studies. Uto (2021) suggested to strengthen ties between research in ATS and testing theory. More research on these and similar strategies is needed to ensure the validity and robustness of ATS systems (Ke and Ng, 2019; Powers *et al.*, 2002; Uto, 2021).

2.2.3.2 Data resources and generalizability

Seeing that most approaches to ATS are supervised, work in ATS has been shaped by the availability constraints of suitable labeled corpora. The number of sufficiently large corpora with gold standard proficiency estimates is limited (Ke and Ng, 2019; Vajjala, 2018), especially for languages other than English. Learner corpora have become an increasingly important resource for ATS that helps to alleviate this limitation (Vajjala, 2018, p. 80). In this context, researchers have started to train ATS models on meta information regarding learners' course levels rather than human expert ratings. Course levels have been argued to be suitable proxies

for L2 proficiency ratings in SLA complexity research, too (Wolfe-Quintero *et al.*, 1998, p. 9), even though they result in less homogeneous characterizations of proficiency than expert ratings (Ortega, 2003, p. 502; Verspoor *et al.*, 2012, pp. 243–244). For underage learners, also age is sometimes used as proxy of proficiency (e.g., De Clercq and Housen, 2017; Xanthos and Gillis, 2010). The modest number of suitable training data also imposes limitations in terms of rating scales. Most available corpora provide holistic proficiency scales but not ratings along relevant sub-dimensions of proficiency (Ke and Ng, 2019), such as grammar control, accuracy, functional adequacy, coherence, or persuasiveness. This restricts work on ATS for sub-dimensions of proficiency which is especially relevant for AWE systems and formative feedback (Uto, 2021, p. 461).

The limited amount of data hinders cross-corpus testing and promotes prompt-specific ATS models because most corpora and data sets elicit language data through a relatively narrow range of prompts (Vajjala, 2018, p. 80). Cross-corpus generalizability is an important performance marker of machine learning models. Predictive models are designed to be applied to unseen data from a predefined application domain to solve a specific task—in this case ATS. To test if a model performs well in one domain, it has to be evaluated on test data that is representative for that domain. Although testing a model on a single test set is a common procedure, it is generally preferred to test it on multiple independent data sets that are representative for the intended application domain. A prerequisite for this is that multiple independent, labeled data sets are available that are representative for an application domain. As discussed above, this is often not the case for ATS, especially for languages other than English.

Cross-prompt and cross-domain generalizability are related constructs and both desirable qualities of machine learning models. Cross-domain generalizability refers to the robust performance of a model on data from another (but typically related) application domain (Maniyar *et al.*, 2020). Cross-prompt (or cross-task) generalizability refers to the robust performance of a model on data from the same application domain but elicited based on another input prompt or task. In ATS, it is in practice most important to be able to use a model for rating texts that were elicited using new prompts (Ke and Ng, 2019; Uto, 2021). This is a particular concern seeing that task contexts are relevant components of many proficiency frameworks (see Section 2.2.1). Furthermore, source materials and tasks have been shown to affect language performance (in terms of the CAF triad, see Section 2.1.3.2). The most common solution to this challenge has been to focus on prompt-specific ATS models which may generalize across corpora based on the same prompt but are not tested—or intended to be used—across prompts

(Phandi *et al.*, 2015; Uto, 2021). In practice, this is a valid—and in fact the preferred (Ke and Ng, 2019)—approach when designing a system for a well-defined and stable application context in which the task prompt remains constant and is used to elicit language samples from many different learners. To support ATS for less well-defined application domains and to address the lack of labeled training data (Uto, 2021), some work has been dedicated to training cross-prompt models (Uto, 2021) and cross-domain testing (Ke and Ng, 2019). For this, several methods have been explored, such as multi-domain learning—that is training on data from multiple domains (Maniyar *et al.*, 2020)—for example using the data from the Automatic Essay Scoring Competition (Kaggle, 2012). Furthermore, researchers have worked on domain adaptation (e.g., Phandi *et al.*, 2015) and the identification of less task-sensitive features (e.g., Zesch *et al.*, 2015). The latter also has valuable implications for research on SLA complexity and writing quality research. Vajjala (2018) explored a multi-corpus study design to compare feature informativeness and feature weights across corpora using different prompts. In contrast to a cross-corpus study, her multi-corpus study compared different prompt-specific models to understand which insights for writing quality research can generalize across prompts and which are domain-specific. However, more research on cross-prompt as well as on cross-domain models for ATS is needed (see also Crossley, 2020).

2.2.3.3 Neural and feature-based approaches

Feature-based machine learning approaches have a long tradition in ATS. However, over the last decade neural network-based approaches have become increasingly popular (for an overview, see Hussein *et al.*, 2019; Uto, 2021). As with other computational linguistic applications, neural machine learning approaches have shown to achieve SOTA results in ATS without requiring explicit feature engineering. The absence of explicit feature encoding has been argued to be both an advantage and a disadvantage. While feature engineering is resource- and time consuming (Hussein *et al.*, 2019; Uto, 2021), it makes the prediction of ATS models more interpretable. Neural approaches yield little insights for research on writing quality and SLA complexity and can be less readily used to inform formative feedback in AWE (Crossley, 2020; Uto, 2021). This continues to be true despite recent efforts to link neural ATS predictions with interprerable text characteristics (e.g., Dong and Zhang, 2016). In practice, this lack of transparency also reduces the applicability of ATS systems in high-stakes assessments. Automatic scoring must be comprehensible, transparent and communicable (Attali and Burstein, 2006, p. 6; Powers *et al.*, 2002, p. 2; Zhang, 2013, p. 13). This not only ensures that the va-

lidity of a system can be properly assessed (Attali and Burstein, 2006, p. 6). It also promotes acceptance of automatic scoring among users, which is important in light of the persistent criticism and skepticism regarding the use of ATS in practice (for details, see Attali and Burstein, 2006; Seifried *et al.*, 2016). Furthermore, neural approaches typically require large amounts of data which are not available for many languages other than English or specific scoring dimensions (Ke and Ng, 2019). Thus, feature-based approaches continue to be relevant for ATS alongside neural approaches.

Feature-based ATS approaches have utilized a broad range of textual features that can be connected to the CAF triad and the complexity domains discussed in Section 2.1.2. Generally, text length features have been found to be successful surface level predictors of proficiency and writing quality (Ke and Ng, 2019; Vajjala, 2018). This is problematic because ATS systems should measure text characteristics that go beyond text length (Attali and Burstein, 2006, pp. 1–2). Text length is uninformative because it can be augmented by any number of linguistic text characteristics, it is construct-irrelevant for writing quality and language proficiency, and it can be easily manipulated in adversarial input (Attali and Burstein, 2006; Quinlan *et al.*, 2009). Measures of accuracy and complexity have proven linguistically more meaningful while also supporting robust and valid ATS. Besides error measures (Borade and Netak, 2020; Ke and Ng, 2019; Vajjala, 2018), measures of syntactic complexity (Borade and Netak, 2020; Crossley, 2020; Crossley *et al.*, 2015, 2016b; Ke and Ng, 2019; Vajjala, 2018), lexical complexity (Borade and Netak, 2020; Chodorow and Burstein, 2004; Crossley, 2020; Crossley *et al.*, 2015, 2016b; Ke and Ng, 2019; Vajjala, 2018), and text cohesion (Crossley, 2020; Crossley *et al.*, 2015; Ke and Ng, 2019; Vajjala, 2018) have been particularly prominent in ATS approaches. Crossley (2020) provides a relatively recent overview of how features of these domains have been used in AES with a focus on writing quality assessment. Some studies explored the use of other discourse complexity measures, including argumentation structure and topic measures (Chodorow and Burstein, 2004; Crossley *et al.*, 2015, 2016b; Ke and Ng, 2019; Vajjala, 2018). Also measures of semantic complexity have been frequently used for ATS, but were mostly reported as aspects of lexical or cohesion measures (Borade and Netak, 2020; Crossley, 2020; Vajjala, 2018). In this context, the comparison of form-based and meaning-based similarities between language samples and source materials has played a relevant role for the assessment of prompt-based language productions. For this, some prompt-based features have been explored for prompt-specific ATS, especially using LSA (for an overview, see Ke and Ng, 2019). Morphological complexity features have been used considerably less often. No-

table exceptions for languages other than English are Vajjala and Lõo (2013, 2014) and the contributions to ATS in this thesis (Weiss and Meurers, 2019a,b, 2021, Section 5.2). Also readability formulas have been occasionally explored for ATS (Ke and Ng, 2019). In contrast, processing complexity measures have played only a minor role in ATS (but see Weiss and Meurers, 2019a,b, 2021, discussed in Section 5.2).

Beyond these text-based measures, some studies indicated that subject-related measures such as other performance estimates (e.g., grades, literacy skills, background knowledge, or cognitive abilities) and demographic information (e.g., age, gender, L1s, L2s) can improve the accuracy of text feature-based ATS systems (e.g., Chodorow and Burstein, 2004; Crossley *et al.*, 2015, 2016b; Vajjala, 2018). This is in line with the relevance of individual characteristics in some of the theoretical proficiency models discussed in Section 2.2.1. Similarly, some task and register-based measures have been proposed to account for task effects (see Section 2.1.3.2) and context-specific expectations that can differ, for example, across disciplines (Crossley, 2020, p. 417). Crossley *et al.* (2016b) argued that the inclusion of such measures also allows to account for the fact that in most contexts there are different potentially successful writing strategies (see also Crossley *et al.*, 2014b) and that L2 proficiency levels do not require a homogeneous profile of competencies (see also Council of Europe, 2001; Hulstijn, 2011). However, these types of measures can only be applied in prompt-specific ATS.

Despite this broad exploration of text features and linguistic domains, there is little consensus on which combination of features best predicts proficiency and text quality in L1 or L2 productions (Attali and Burstein, 2006; Crossley, 2020; Vajjala, 2018, p. 80). This is due to the lack of feature-based models focusing on cross-prompt scoring (see previous discussion). Generally, though, accuracy as well as syntactic, lexical, and discourse complexity seem to be systematically among the most important measures (Attali, 2007; Crossley, 2020; Vajjala, 2018) and a diverse set of features seems to outperform homogeneous feature sets (e.g., Vajjala, 2018).

2.3 Automatic readability assessment

Automatic Readability Assessment (ARA) seeks to align texts with the reading skills of their prospective readers by predicting the comprehensibility of a text for a pre-defined target group, such as L2 learners or children reading in their L1 (e.g., Glöckner *et al.*, 2006; Vajjala, 2022). The comprehensibility of a text has been shown to be influenced by the interaction

between text characteristics, reader characteristics, and reading goal(s) (see Section 2.3.1.1). This makes readability assessment an inherently interdisciplinary endeavor (Benjamin, 2012; Collins-Thompson, 2014) which has been pursued for a variety of application domains since the early 20th century. The following section provides a focused background on automatic approaches to readability assessment using computational linguistic methods. After a brief sketch outlining the general concept of text readability (Section 2.3.1), I will describe the main application domains of ARA (Section 2.3.2) before discussing the central methods, trends and challenges in computational linguistic research on the topic (Section 2.3.3). For a discussion of ARA for German, please see the structured literature survey in Section 4.3.

2.3.1 What is text readability?

Text readability is a broad term that Dale and Chall (1949) generally defined as the result of all components contributing to or impeding text understanding, reading fluency, and interest while reading. Three components have been identified as particularly relevant for readability across research disciplines: 'text characteristics' and 'reader characteristics' (Collins-Thompson, 2014; Long *et al.*, 2006; Smith *et al.*, 2021; Vajjala, 2022; Yin, 1985; Zwaan and Rapp, 2006) as well as 'reading goal(s)' (Valencia *et al.*, 2014; Zwaan and Rapp, 2006). A text can be considered readable for a given reader if it allows them to build a sufficiently accurate mental representation of the text which they can integrate with their prior knowledge (see Section 2.3.1.1) in a way that fosters new inferences (Long *et al.*, 2006, pp. 824–825). In psychological and psycho-linguistic research this has been referred to as a 'coherent discourse model' (Long *et al.*, 2006, p. 825) or 'situation model' (Zwaan and Rapp, 2006, p. 727). The reading goal (e.g., learning or pleasure) impacts how readers build their discourse model and which reading strategies they employ (Zwaan and Rapp, 2006, p. 729). Readers differ in their ability to from a coherent discourse model for a specific text and reading goal. The sophistication and accuracy of an individual readers' discourse model can vary based on text characteristics—such as legibility, linguistic properties, and extra-linguistic materials—and reader characteristics—such as working memory capacity and prior knowledge—(Zwaan and Rapp, 2006, p. 727). Most texts vary in the degree of their comprehensibility. The extremes of perfectly comprehensible (i.e., a maximally sophisticated and accurate discourse model can be formed and can be highly interconnected with a reader's prior knowledge) or incomprehensible texts (i.e., no or only a simple and inaccurate discourse model can be formed without connections to prior knowledge) are not the norm. Text generally fall on a continuum between

these two poles. Hence, the question is not whether a text is readable or not, but whether it is *sufficiently* readable for the specified reading goal(s).

For ARA, we seek to identify texts for which we can assume that the discourse model is sufficiently sophisticated and accurate to support the intended reading goal. At the same time, ARA may seek to optimize different variables beyond reading comprehension depending on the reading goal. For example, when identifying reading materials for language learners, an ideal alignment between reader characteristics and text characteristics should provide challenging input to readers in their Zone of Proximal Development (ZPD) (Vygotsky, 1978), prioritizing language learning outcomes over a maximally coherent discourse model. Other relevant variables ARA may seek to optimize are pleasure or reading speed.[4] In short, ARA ultimately is an alignment task between reader and text for a given reading goal (for a similar view, see also Beinborn *et al.*, 2012). When conceptualized this way, it becomes evident that unlike many other established computational linguistic tasks—such as topic prediction or machine translation—ARA is intrinsically "user- or population-specific" (Collins-Thompson, 2014, p. 104).

2.3.1.1 The role of text characteristics, reader characteristics, and reading goals

ARA research has historically focused on text characteristics (Collins-Thompson, 2014; DuBay, 2004; Vajjala, 2022), even though researchers have acknowledged that reader characteristics such as motivation and prior knowledge play an important role for comprehension (e.g., Bailin and Grafstein, 2001; Collins-Thompson, 2014). Text characteristics include layout factors (such as font size and color or contrast) as well as language and content factors (such as coherence, syntax, lexicon, semantics). To better separate them, researchers often distinguish 'legibility'—influenced by layout factors—from 'readability'—influenced by language and content—(e.g., Dale and Chall, 1949; DuBay, 2004). ARA focuses near exclusively on readability in this narrow sense, mostly ignoring questions of legibility (Collins-Thompson, 2014; Vajjala, 2022). Features of linguistic complexity have played a central role for this text-based view on readability. There has also been a distinct focus on measures of discourse cohesion in psychological and psycho-linguistic research on discourse comprehension (McCarthy and McNamara, 2021; McNamara and Kintsch, 1996; McNamara *et al.*, 1996, 2011; Ozuru *et al.*, 2009; Smith *et al.*, 2021). I discuss this in more detail in Section 2.3.3.3.

[4] I described scenarios that focus on a single reading goal. However, in practice readers and teachers can have multiple (primary or secondary) reading goals.

Reader characteristics have been mostly researched in psycho-linguistics and psychology (McCarthy and McNamara, 2021; Smith *et al.*, 2021). These include group characteristics and individual characteristics. In ARA research, reader characteristics have mostly been considered in terms of group properties (Collins-Thompson, 2014), aiming to provide population-specific ARA models (for a notable exception, see Chen and Meurers, 2019). Most attention has been paid to the population differences between young L1 readers, who still need to mature their literacy skills, and literate adult L2 readers (Collins-Thompson, 2014; Heilman *et al.*, 2007; Sung *et al.*, 2015; Xia *et al.*, 2016). To illustrate the central differences between these two groups and their implications for text readability, let us revisit the notions of BICS and CALP proposed by Cummins (2000) and discussed in Section 2.2.1, p. 44. Young L1 readers are typically beginning their literacy acquisition in a formal educational setting after having developed advanced BICS but prior to the acquisition of CALP. Thus, they may struggle with domain-specific or academic vocabulary and lack advanced reading strategies, but they have generally mastered the common morpho-syntactic constructs and lexical items of the language in which they are reading. Adult L2 learners are often already highly literate in their L1(s) and have an advanced CALP in at least one other language. They should have mastered central reading strategies and may have higher CALP in the language they are reading due to transfer effects from their L1(s) and other L2s. However, they are prone to struggle with common morpho-syntactic constructs and lexical items of the language in which they are reading. Hence, these two different target groups will experience different concepts and linguistic constructs as challenging (Collins-Thompson, 2014, pp. 114–115) which necessarily impacts the alignment between text and reader. For an attempt to generalize an L1 ARA model to L2 contexts, see Xia *et al.* (2016), which I discuss in more detail in Section 2.3.3.1 (p. 73).

The readability of a text for a specific reader is further determined by a reader's individual properties. These can be situational or stable. Situational individual properties can vary across reading contexts and include, for example, learners' motivation, current level of stress, or interest. Stable individual properties do not vary across contexts although they might change over time. Examples are language proficiency, working memory, degree of literacy, or prior knowledge. Individual properties have a strong impact on reading comprehension which goes beyond language proficiency (Yin, 1985). They can can lead to substantial inter-individual differences in text comprehension, especially with regard to readers' ability to interpret and make higher-level inferences based on a text (Long *et al.*, 2006, pp. 801–802). Long *et al.* (2006, p. 802) identified five central stable individual properties that impact reading compre-

hension in literate, adult L1 readers: word identification skills, working memory capacity, the ability to inhibit irrelevant information and signals, print exposure, and prior knowledge (for a similar view, see also Smith *et al.*, 2021).

I will focus on the role of prior knowledge for reading comprehension because it has received the most attention in psychological research on reading comprehension (e.g., Best *et al.*, 2006, 2008; Kamalski *et al.*, 2008; Lipson, 1982; McCarthy and McNamara, 2021; Ozuru *et al.*, 2009; Smith *et al.*, 2021; Yin, 1985; Zwaan and Rapp, 2006). Prior knowledge is a general notion that includes different types of knowledge. It encompasses general world knowledge (Smith *et al.*, 2021, p. 216) but also language- or culture-specific pragmatic communicative knowledge (Yin, 1985, p. 376). Both serve as an important frame of reference for readers to correctly interpret a text. However, most research on readability has focused on the role of domain-specific knowledge (e.g., Kamalski *et al.*, 2008; McCarthy *et al.*, 2018; McNamara *et al.*, 1996, 2011; Ozuru *et al.*, 2009; Smith *et al.*, 2021). Prior domain-specific knowledge has been found to benefit readers' text comprehension irrespective of their reading skills (Lipson, 1982; McCarthy and McNamara, 2021; McNamara *et al.*, 2011; Smith *et al.*, 2021). Smith *et al.* (2021, pp. 226–227) found in their survey of studies on the role of prior knowledge for reading comprehension in children, that children can to a certain degree compensate for a lack of prior knowledge with high literacy skills and vice versa. McCarthy and McNamara (2021, p. 196) reported that prior knowledge can account for 30%–60% of variance in reading comprehension. However, the role of prior knowledge varies with age and genre (see discussion in Smith *et al.*, 2021, p. 219). The beneficial effect of prior knowledge is mediated by its quality, specifically its accessibility—How easily can it be retrieved from long-term memory?—and its accuracy—Is the knowledge factually correct?—(Smith *et al.*, 2021, p. 217). For example, Lipson (1982) found that prior knowledge made readers more resilient towards the integration of new information that contradicted their (incorrect) prior knowledge in the sense that readers prioritized their prior knowledge over the textual information regardless of the factual correctness of the prior knowledge. McCarthy and McNamara (2021) proposed to account for some of this variance by introducing the 'Multidimensional Knowledge in Text Comprehension' framework. It differentiates declarative content knowledge along the dimensions of amount, accuracy, specificity, and coherence of prior knowledge.

Special attention has been paid to the interaction between prior knowledge, literacy, and text cohesion. High cohesion benefits discourse comprehension for readers with little domain-specific knowledge (Smith *et al.*, 2021, pp. 218–219). High textual cohesion makes connec-

tions more salient and reduces the cognitive load of building a well-connected mental representation of a text (McCarthy and McNamara, 2021; Smith *et al.*, 2021). It can thus help to compensate for unavailable domain knowledge. Also, the effort required to form connections between prior knowledge and the situation model depends on the accessibility of prior knowledge. The more accessible prior knowledge is, the less effort readers' have when connecting it with their situation model (for a similar argument, see Smith *et al.*, 2021, p. 218). High textual cohesion can help to reduce the retrieval costs of less accessible prior knowledge because it makes the connections between prior knowledge and text contents more salient (Smith *et al.*, 2021, p. 219). In contrast, readers with high domain-specific knowledge can benefit from less cohesive texts (McNamara and Kintsch, 1996; McNamara *et al.*, 1996, 2011; Smith *et al.*, 2021). It has been argued that low cohesion texts result in a better connected and more sophisticated mental text representation in readers with high domain-specific knowledge because their higher cognitive demands foster cognitive activation during reading (Kamalski *et al.*, 2008; McNamara and Kintsch, 1996; McNamara *et al.*, 1996; Smith *et al.*, 2021). This effect is known as the reverse cohesion effect (see Section 2.1.3.1, p. 33). Smith *et al.* (2021, p. 228) refer to it as a specific type of expertise reversal effect based on which heavy scaffolding benefits beginning learners but should be reduced for more advanced learners. However, the reverse cohesion effect has been argued to be absent in readers with high reading skills (Ozuru *et al.*, 2009) and could not be reproduced for persuasive texts (Kamalski *et al.*, 2008).

Most research on readability has focused on text characteristics and reader characteristics, even though reading goals have been acknowledged as a relevant component (Zwaan and Rapp, 2006). Readers can pursue different reading goals, such as content-matter or language learning, information retrieval, identifying procedural instructions, or entertainment. Readers can also pursue different primary and secondary reading goals, for example, entertainment and information retrieval. The role of reading goals for text comprehension has also been addressed indirectly through work on genre effects. The most commonly studied genres in readability research are narrative texts, factual or persuasive expository texts, and procedural texts—such as manuals or recipes—(Zwaan and Rapp, 2006, p. 728). Although the purpose of reading and the genre of text are clearly different, in practice they are often directly related. For example, narrative texts are more often read for pleasure than procedural instructions. Researchers have found that discourse genre plays a central role for reading comprehension (Best *et al.*, 2008; Kamalski *et al.*, 2008; McNamara *et al.*, 2011; Zwaan and Rapp, 2006). Genre has shown to influence the role of prior knowledge for discourse comprehension (e.g., Kamalski

et al., 2008; Smith *et al.*, 2021) and to influence the cognitive processes and mental representations associated with reading (for details, see Zwaan and Rapp, 2006, p. 729). According to the survey by Smith *et al.* (2021), studies systematically found that the effect of prior knowledge on reading comprehension in children was mediated by text genre. Specifically, prior knowledge played less of a role for narrative than expository texts. Potential explanations for this are that children are more practiced in the comprehension of narratives (Best *et al.*, 2008; Smith *et al.*, 2021) and that expository texts require more references to prior knowledge than narratives (Smith *et al.*, 2021). More work is still needed to disentangle effects of discourse genre and reading goal on discourse comprehension.

2.3.1.2 Frameworks of discourse comprehension: the construction-integration model

Several frameworks have been suggested to formally account for the interaction between text characteristics and reader characteristics. One of the most influential frameworks has been Kintsch's 'construction-integration model' (Kintsch, 1988; Wharton and Kintsch, 1991). Kintsch (1988) identified it as one of the first models of discourse comprehension that conceptualized comprehension as a bottom-up process (starting with decoding word meanings) rather than a top-down process (initially guided by readers' prior knowledge). The model distinguishes two comprehension stages: the construction stage and the integration stage. The construction stage describes the bottom-up process of creating a general associative mental network of concepts and propositions. This network is referred to as the 'text base'. The text base is created in a four-step process in which readers are assumed to decode the concepts and propositions explicitly encoded in the text and to retrieve, activate, and connect all elements related to these concepts and propositions in their network of prior knowledge (Kintsch, 1988, p. 166). The construction stage is assumed to rely only on readers' morpho-syntactic and lexico-semantic knowledge as well as the general world knowledge needed to form text-based network of concepts and propositions. At this point, readers have constructed an incoherent and inconsistent text base. It lacks relevant connections and contains incomplete or inappropriate concepts due to two main reasons. First, the construction stage takes place on-line and is thus based on incomplete information (Kintsch, 1988, p. 166). Second, the construction stage does not utilize readers' discourse knowledge to prune or emphasize nodes and connections based on their relevance (Kintsch, 1988, p. 168). This issue is addressed in the integration stage. It enriches the text base with additional concepts and connections between nodes but also inactivates and prunes irrelevant connections and nodes to create a coherent 'situation

model' (Kintsch, 1988, p. 164). To do so, readers use their socio-pragmatic knowledge about the language they are reading, their prior domain-specific knowledge, their personal experiences and beliefs, and their general world knowledge. It is assumed that the construction stage and the integration stage alternate repeatedly during discourse processing. After a text base for a phrase or sentence has been created, it is edited into a situation model in the integration phase. The two networks (text base and situation model) are gradually expanded during reading (or listening) based on the new input (Kintsch, 1988, p. 168).

Text base and situation model can vary in their quality (McNamara and Kintsch, 1996, p. 252). A reader's text base may be incomplete or inaccurate due to poor text quality or an insufficient decoding process, for example because of inattentive reading or deficits in word identification (McNamara and Kintsch, 1996; Smith *et al.*, 2021). A reader's situation model may lack connections, activate irrelevant concepts, or include incorrect concepts or connections because of insufficient or incorrect prior knowledge or a poor application thereof (McNamara and Kintsch, 1996; Smith *et al.*, 2021). A poorly elaborated situation model is close to the text base. However, readers can also considerably extend the situation model beyond the text base and overwrite or disable connections from the text base. This is particularly likely to happen to readers with high background knowledge and a poor text base (McNamara and Kintsch, 1996; Smith *et al.*, 2021). Thus, the correspondence between text base and situation model may vary greatly (McNamara and Kintsch, 1996, pp. 252–253) and depends considerably on readers' prior knowledge. Different discourse comprehension test items have been proposed to estimate the quality of text base and situation model independently (see McCarthy and McNamara, 2021, p. 199). When comparing text base and situation model it is also important to note that both are stored differently. The text base is stored in readers' in working memory whereas the situation model is committed to long-term memory (Smith *et al.*, 2021, p. 215). Thus, working memory capacity also plays a relevant role in the construction-integration model, albeit less pronounced than prior knowledge. The better the situation model is inter-connected with readers' prior knowledge, the more of it is retained in long-term memory and integrated into readers' network of background knowledge (McNamara and Kintsch, 1996, p. 253). However, when the working memory capacity is overloaded in discourse comprehension, this can hinder the integration of the text base into the situation model as well as the enrichment of the situation model with further connections. The effort required to form connections between prior knowledge and the situation model depends on the accessible prior knowledge in the sense that accessible prior knowledge reduces readers' effort

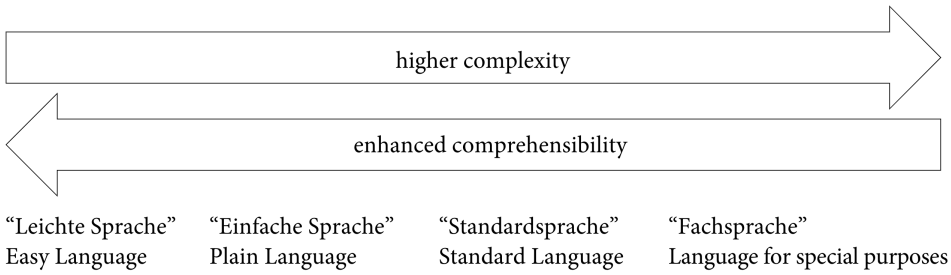


Figure 2.2: *Assumed inverse, continuous relationship between complexity and comprehensibility illustrated for German language varieties for different target groups from Hansen-Schirra and Maaß (2020, Figure 1, p. 18)*

(see Section 2.3.1.1). High textual cohesion can help to reduce the cognitive load required to form these connections if prior knowledge is unavailable or inaccessible (Smith *et al.*, 2021, p. 219).

Terminological note In computational linguistic work on ARA, the notions ‘text complexity’ and ‘text readability’ are often used interchangeably (e.g., Vajjala, 2022). This is based on the implicit assumption that texts can be arranged on a continuous scale between two opposite poles of ‘simple’ to ‘complex’ and that the comprehensibility of a text is indicated by its position on this scale. This is illustrated in Figure 2.2 (originally from Hansen-Schirra and Maaß, 2020, p. 18). Hansen-Schirra and Maaß (2020) located simplified and non-simplified language varieties on a continuous scale of language complexity and comprehensibility. Although this assumption seems plausible at first (and may empirically hold for some target reader groups and reading goals), it confounds two separate concepts. As discussed in Section 2.3.1, text characteristics are one of several factors influencing discourse comprehension. Also, higher linguistic complexity is not always detrimental to text comprehension. The reverse cohesion effect is an example of when lower complexity (in the sense of more explicit, cohesive, and cognitively less demanding) does not benefit reading comprehension for certain target populations (see Section 2.1.3.1, p. 2.1.3.1). Another example are L2 readers with high CALP, who are less likely to struggle with high demands on their CALP than high demands on their BICS (see Section 2.2.1, p. 44). Using the notion of text complexity synonymous to the notion of readability invites the misconception that less complex texts are always more comprehensible, thus falling into the same fallacy as referring to more complex L2 writing as more proficient or advanced (see Section 2.1.1.2, p. 14). I argue against using both terms interchangeably.

2.3.2 Application domains

ARA has been utilized in various application contexts. This section discusses common uses for ARA illustrated with exemplary studies (for a systematic review focused on German, see Section 4.3). I propose to structure uses of ARA into three domains based on the goal they try to optimize by controlling text readability through ARA: i) education and language learning (optimizing learning outcomes), ii) accessibility and information retrieval (optimizing text comprehension), and iii) user experience and quality control (optimizing a task that is being mediated by text comprehension).

2.3.2.1 Education and language learning

Education, both inside and outside of instructed settings, is one of the most common application domains for ARA (Benjamin, 2012; Collins-Thompson, 2014). Readability scores have been used to support multiple stakeholders in the language learning field alike. This includes learners and teachers, who need to select appropriate learning materials, as well as education content providers and publishers, who create learning and teaching materials such as schoolbooks. In this context, special attention has been paid to provide learners with input in their ZPD (Vygotsky, 1978) which challenges them at their current level of language competence. Both overchallenge and underchallenge can negatively impact learning outcomes (Sung *et al.*, 2015, p. 372). Readability also plays an important role for the assessment of item difficulty for content-matter (e.g., Höttecke *et al.*, 2018) and reading tests (e.g., Ludewig *et al.*, 2022). Research on ARA has predominantly focused on the identification of leveled reading materials for pupils in schools (Collins-Thompson, 2014), even though there have also been early approaches focusing on ARA for L2 readers (e.g., Crossley *et al.*, 2008; Heilman *et al.*, 2007; Schwarm and Ostendorf, 2005). Most ARA studies for (language) learning focus on providing new SOTA models to predict readability (e.g., Feng *et al.*, 2010; François and Fairon, 2012; Imperial and Ong, 2021; Lee and Lee, 2020; Naderi *et al.*, 2019a; Saddiki *et al.*, 2018; Todirascu *et al.*, 2013; Vajjala and Meurers, 2012). However, less researchers have also incorporated their models into publicly available systems where users without skills in programming or machine learning can access them (Benjamin, 2012). In the following, I focus on studies that made their ARA models accessible.

Pilán *et al.* (2016) proposed a feature-based readability classifier for Swedish L2 learners which predicts readability on the CEFR scale (A1–C1) for entire documents and individual

sentences. They trained their models on reading passages and example sentences from leveled books for Swedish language learners. Pilán *et al.* (2016) reported an accuracy of 81.3% (adjacent accuracy = 97.0%) for document-level and 63.4% (adjacent accuracy = 92.0%) for sentence-level classification in 10-folds cross-validation (10-CV). In contrast, the Läsbarthetsindex (LIX) by Björnsson (1983) did not substantially improve over the majority baseline for either classification. Pilán *et al.* (2016) integrated their readability models into the online language learning platform Lärka (`https://spraakbanken.gu.se/larka/`). The Lärka web platform targets researchers, teachers, and language learners alike. It provides corpus-based exercises for language learning and teaching as well as data and an annotation editor for researchers. Similarly, Sung *et al.* (2015) proposed a Chinese L2 readability classifier that predicts the readability of texts on the CEFR scale (A1–C2) which achieved an accuracy of 75.0% (adjacent accuracy = 99.6%). They compiled a corpus of reading materials used in Chinese L2 classrooms which they asked five experienced Chinese L2 teachers to rate resolving disagreements through discussion. Similar to Vajjala and Meurers (2012), Sung *et al.* (2015) focused on features associated with absolute complexity or relative complexity (see Section 2.1.1.2, p. 12) for Chinese L2 learners to inform their classifier. They incorporated their model into a web platform (`www.chinesereadability.net`) that allows users to obtain readability scores for their input texts. Importantly, they utilize the linguistic insights provided from their feature-based approach to offer a diagnostic function for users. The system identifies and visualizes the linguistic constructs and their positions in the input text that users need to alter to obtain a different readability score. Deconstructing readability into individual linguistic domains is not only desirable for text adaptation. Beinborn *et al.* (2012) argued that breaking down text readability along relevant linguistic domains can also help learners to identify competence-adaptive materials for self-directed learning. They argued that estimates along individual linguistic dimensions are more informative and better account for heterogeneous competence profiles than holistic readability scores.

ARA models have not only been incorporated into web platforms for the analysis of individual texts, but also been used as components in tutoring systems. The connection of ARA models with tutoring systems that model learners' individual properties allows to provide a more individual alignment between users and text properties. However, most tutoring systems utilize simple ARA methods to identify authentic reading materials at learners' level of proficiency rather than SOTA ARA models. For example, the REAP tutoring system (Brown and Eskenazi, 2004; Heilman *et al.*, 2010) supports learners' individualized English

L2 vocabulary training through reading exercises (for a Portuguese version of the system, see Marujo *et al.*, 2009). It identifies authentic web materials matching readers' proficiency level through a predominantly vocabulary-based model of learners' reading skills. The Read & Improve reading tutoring system prototype (Watson and Kochmar, 2021) uses a more sophisticated approach to readability assessment. It provides English L2 learners with recent news at a reading level that matches learners' current proficiency level. Learners' proficiency is identified through a learner model that is informed by the Write & Improve platform (`https://writeandimprove.com/`) which uses ATS to provide learners with formative writing feedback while holistically assessing their level of proficiency. Both web platforms are linked through a shared user account. The readability of texts is identified using a feature-based machine learning model that ranks texts in terms of their readability. It uses the features proposed by Xia *et al.* (2016) and was trained specifically on leveled news data for English L2 and L1 learners. To foster vocabulary learning, the Read & Improve system links each word in a reading text to different dictionary definitions and a co-occurrence word cloud. The system further allows users to submit summaries of the texts that they read to test their writing proficiency and reading comprehension. Finally, users can view their reading and writing development through a separate panel as a form of longitudinal feedback.

2.3.2.2 (Web) accessibility and information retrieval

A second common application domain of ARA is concerned with (web) accessibility. This line of research focuses on the identification of accessible materials often for the purpose of information retrieval and typically analyzing web materials (Collins-Thompson, 2014). While there has been some work on ARA in general purpose information retrieval systems (e.g., Kim *et al.*, 2012; Pera and Ng, 2012; Russell, 2011), most work in this area has focused on readers with low literacy skills or special communication needs, for example neuro-atypical readers such as dyslexic readers (Rello *et al.*, 2012, 2013a,b; Sitbon and Bellot, 2008), readers with an Autism Spectrum Disorder (Eraslan *et al.*, 2017, 2021; Yaneva *et al.*, 2015) or readers with cognitive disabilities (Abedi *et al.*, 2012; Feng *et al.*, 2009).[5] As Collins-Thompson (2014) observed, this application domain is also closely linked to work on text simplification (e.g., Bingel *et al.*, 2018; Rello *et al.*, 2013c; Yaneva *et al.*, 2016), see also Siddharthan (2014); Štajner (2021).

[5]Note that 'neuro-atypical readers' is a catch-all phrase commonly used in work on ARA for accessibility, but does not define a homogeneous target group with shared reading needs.

Readability scores may be used to help readers with special needs to find texts that align with their degree of reading competence or to inform content providers (such as news outlets or government information channels) about the accessibility of their materials. This includes work on the validation of reading materials that need to comply with specific guidelines for accessible language, such as Yaneva's (2015) approach for the validation of English Easy-to-read (Freyhoff *et al.*, 1998; Nomura *et al.*, 2010) materials. An example for a fully implemented system that supports the retrieval of accessible reading materials is the Read-X system (Miltsakaki and Troutt, 2008). It is an English search engine that retrieves reading materials for low literate adolescents and adults and classifies them based on their topic and readability. The system measures readability using three simple readability formulas for English L1 readers: the Automatic Readability Index (ARI, Amstad, 1978), the LIX (Björnsson, 1983), and the Coleman-Liau index (Coleman and Liau, 1975). As a second predictor of readability, the system uses domain-specific frequency lists to account for the fact that domain-specific prior knowledge greatly influences readability (Miltsakaki, 2009). Vocabulary that the model predicts to be unknown to readers is highlighted and linked to WordNet as an external information resource. This prediction is based on information that users provided regarding their educational background and familiarity with the topic at hand. Similar systems are the LAWSE search engine (Ott and Meurers, 2011) or the FLAIR search engine (`https://flair.schule/`; Chinkina *et al.*, 2016) for German and English L2 learners. Many systems such as Read-X, LAWSE, and FLAIR level the retrieved materials with traditional readability formulas for L1 readers instead of target population-specific ARA methods (for a notable exception, see Collins-Thompson *et al.*, 2011). The KANSAS search engine (`www.kansas-suche.de`; Dittrich *et al.*, 2019; Weiss *et al.*, 2018) provides readability estimates using a customized readability formula for low literate adolescents and adults in Germany. The formula was closely based on the standards used in practice to assess reading skills in low literate adults in Germany and group them into appropriate course levels in adult education centers (for details, see Weiss *et al.*, 2018). As such, it is one of the few text retrieval systems that is based on the available specific guidelines for accessible language for the intended target group rather than a general purpose readability formula. To ensure that teachers can always retrieve accessible reading materials even for low levels of literacy, the KANSAS search engine is a hybrid system that allows to query a pre-compiled corpus of leveled reading materials or perform a web search (Dittrich *et al.*, 2019).

ARA for (web) accessibility does not exclusively target special-needs readers. There has

been considerable work on the accessibility of legal or municipal texts for non-experts (Ojha *et al.*, 2018; vor der Brück *et al.*, 2008), on the readability of privacy policies (Sunyaev *et al.*, 2015) and user manuals (Andersson and Szewczyk, 2011), and on the accessibility of physical and mental health care information (King *et al.*, 2003; Misra *et al.*, 2013; Paul *et al.*, 2021; Skierkowski *et al.*, 2019). For example, Kiwanuka *et al.* (2017) studied the readability of English patient information resources for gender affirmative surgery on the web which they retrieved using four search terms related to different forms of gender affirmative surgeries. They assessed readability using ten readability formulas returning grade level estimates as reading levels. Across readability formulas and search terms, they obtained an average grade level of 13.4 (corresponding to early university-level reading skills) after confirming the comparability of the readability formulas' estimates. This lies well above the recommended reading level set at 6th grade reading skills by the National Institute of Health and American Medical Association. These findings are corroborated by Vargas *et al.* (2017) who study the readability and quality of resources for gender affirming surgery found on the web using the search term "transgender surgery". Using the same ARA set-up as Kiwanuka *et al.* (2017), they obtained an average grade level of 14.7. Their additional manual assessment of resource quality using two expert raters found a non-linear, predominantly negative correlation between readability and quality.

2.3.2.3 User experience and quality control

Readability has also been identified as an important measure of user experience and quality control. One important application domain for ARA is the evaluation of NLP systems linked to Natural Language Generation (NLG). Most prominently, ARA has been used in work on automatic text simplification as one of several measures to quantify the degree of simplification achieved by a text simplification system (Siddharthan, 2014; Štajner, 2021). Beyond this, ARA has also been used in machine translation to create leveled translations that align to different target audiences (e.g., Agrawal and Carpuat, 2019; Marchisio *et al.*, 2019; Stymne *et al.*, 2013). Similarly, readability or comprehensibility have been recommended as a design factor for conversational agents (CA) because mismatches between users' skills and the CA' language use can be detrimental to user experience or even impede the communicative goals of the interaction (Gnewuch *et al.*, 2018; Langevin *et al.*, 2021; Santhanam *et al.*, 2020). ARA has also shown to predict attention on online platforms (Guerini *et al.*, 2012; Pancer *et al.*, 2019) and was linked to user experience and compliance with safety measures (Andersson

and Szewczyk, 2011). Santos *et al.* (2020) used readability features to detect fake news in Brazilian Portuguese. They reached an accuracy of up to 92% based solely on features of text readability. The most informative features included pronoun diversity, LSA-based measures, and the readability formula by Brunet (1978). Santos *et al.* (2020) also outperformed the SOTA results for fake news detection in Brazilian Portuguese by combining readability features with a more traditional approach to fake news detection by Monteiro *et al.* (2018).

2.3.3 Current methods and trends

ARA dates back more than a century to work on traditional readability formulas, see DuBay (2004, 2006) for an overview. Computational linguistic work on ARA using NLP and machine learning has started to emerge in the early 2000s (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Si and Callan, 2001), and became the standard in computational linguistic research on ARA (Collins-Thompson, 2014; Vajjala, 2022). ARA has predominantly and most systematically focused on English (Collins-Thompson, 2014; Vajjala, 2022) even though there has been scattered work on a broad range of languages, such as Arabic (Saddiki *et al.*, 2018), Swedish (Pilán *et al.*, 2016), Chinese (Sung *et al.*, 2015), French (François and Fairon, 2012), Italian (Dell'Orletta *et al.*, 2011), and German (this thesis; Brück and Leveling, 2007). In the following, I discuss the current methods and trends in supervised ARA in terms of machine learning tasks and evaluation metrics (Section 2.3.3.1) as well as rating scales and corpora (Section 2.3.3.2) used. I then compare neural and feature-based approaches and discuss the type of linguistic features used for ARA (Section 2.3.3.3).

2.3.3.1 Machine learning tasks and model evaluation

Most research on ARA focuses on supervised machine learning (Collins-Thompson, 2014; Vajjala, 2022), for notable exceptions see Jameel and Qian (2012); Jameel *et al.* (2012); Martinc *et al.* (2021). In this context, ARA has been predominantly approached as one of three machine learning tasks: (ordinal) classification, regression, and (pair-wise) ranking (Collins-Thompson, 2014; Vajjala, 2022). It has long been debated which of the three approaches is more suitable for readability detection (Collins-Thompson, 2014). However, it has been shown that the choice of linguistic features has a greater impact on model performance than the choice of machine learning algorithm (Kate *et al.*, 2010).

Unlike ranking, classification and regression both assign absolute readability labels which

makes them applicable in very similar contexts. In practice, the choice of algorithm type is closely linked to the properties of the underlying readability scale. The rational behind this choice is similar to the previously discussed considerations for ATS in Section 2.2.3.1 (p. 51). Readability scales often use ordinal or discrete numerical data with a limited range (such as the CEFR scale or a 5-point Likert scale). These data are typically ordinal in the sense that the equidistance of adjacent categories is not guaranteed. Thus, (ordinal) classification is considered as a suitable alternative to regression algorithms. The evaluation metrics associated with ARA naturally depend on the classification algorithm used. Classification is typically evaluated in terms of accuracy or precision, recall, and F1-score on a test set (Collins-Thompson, 2014; Vajjala, 2022). As with ATS, many ARA studies calculate the adjacent accuracy of their models to account for the ordinal nature of the reading scales (Sung *et al.*, 2015, p. 382). For regression models, researchers typically use RMSE, MAE, or Pearson correlation if the predicted labels and the evaluation data use the same reading scale. Spearman rank correlation is also commonly used if the predicted labels are on a different reading scale than the evaluation data (Collins-Thompson, 2014), e.g., in cross-corpus testing studies (see Section 2.3.3.2).

In contrast to regression and classification, ranking approaches assign a relative score based on the available input options. This makes it particularly suitable for tasks where the readability of a pre-defined set of text options should be compared to each other (e.g., to evaluate the success of text simplification by comparing simplified and original versions of texts). As Xia *et al.* (2016) pointed out, this is conceptually accounting for the fact that readability might be better described as a relative than an absolute property of texts in the sense that one text can be more or less readable than another. The lack of an absolute label can also facilitate the generalizability of a model to different reading scales and target populations. For example, Xia *et al.* (2016) experimented with treating the difference between L1 readers and L2 readers as a domain-adaptation problem to address the lack of L2 corpora for ARA (see Section 2.3.3.2). They trained a classification and a ranking model on leveled English L1 data before exploring several strategies to transfer their L1 models to L2 data, including domain adaptation and self learning. They found that while ranking did not yield the best within-domain performance, it could be successfully adapted to L2 data using self learning. However, ranking approaches are also less specific than classification or regression models and are thus not suitable for all application contexts. Also, ranking approaches often lack a reference to prospective readers' language skills because they construct readability primarily as a property between texts, not between a text and a reader. Ranking approaches are typically evaluated in pair-wise compar-

isons by calculating the percentage of correctly ranked text pairs.This is known as pairwise (ranking) accuracy. Pairwise accuracy evaluates only the order of ordinal labels. It is generally less informative than correlations or metrics such as accuracy or f-score because it ignores central aspects of prediction quality. Importantly, unlike Spearman rank correlation, pairwise accuracy ignores the distance between labels (i.e., how much simpler is one text than another?). Also, unlike Pearson correlation or accuracy, it is agnostic to the precise position of documents on a reading scale. Hence, pairwise accuracy is not an ideal metric for all use cases and is not directly comparable to accuracy because the criterion for pairwise accuracy is easier to satisfy than the criterion for accuracy (see also Xia *et al.*, 2016).

Metrics such as accuracy, f-score, or RMSE account for the robustness of a model, that is, how well the model approximates the labels in the test data used as gold standard (see Figure 2.1, p. 53). However, little research has been dedicated to the assessment of the construct validity of ARA models (see discussion in Section 2.2.3.1, pp. 53–54). Vajjala (2022) identified the need for systematic validation of models as a major research desideratum in her survey on computational linguistic research on ARA (for a similar argument, see Collins-Thompson, 2014). A suitable but resource intensive method to test the validity of models are reading experiments that test if predicted readability rankings are in line with empirical observations of reading performance for the target population—such as reading times, performance in reading comprehension tests, or human readability judgments—(Benjamin, 2012; Miltsakaki and Troutt, 2008; Vajjala, 2022). As a less resource-intensive alternative, Vajjala (2022) also proposed to focus on the extrinsic evaluation of ARA models by integrating them into real-world systems (e.g., tutoring systems or search engines, see Section 2.3.2). She reasons that if an ARA model can be successfully used in practice in a system with real-life users, this demonstrates its ecological validity even though it does not provide insights into the construct validity of the predictions. However, seeing that ARA in practice near-exclusively relies on traditional readability formulas rather than SOTA machine learning models (see discussion in Section 2.3.4), we have little insights on the external validity of ARA models.

2.3.3.2 Data resources and generalizability

Research on ARA has been substantially shaped by the availability constraints of labeled corpora that can be used to train supervised machine learning models. Only a limited number of suitable corpora is available for research (Collins-Thompson, 2014; Vajjala, 2022; Xia *et al.*, 2016), this holds especially for languages other than English and for L2 readability corpora

(Xia *et al.*, 2016). Publishers of education materials have been an important resource of leveled data for readability corpora (Vajjala, 2022, pp. 3–5). Textbooks are usually produced for learners at a specific proficiency level. They may even cover multiple proficiency levels as they are designed to be used over an extended period of time in which learners advance in proficiency. Also, textbooks are available for a variety of languages. Thus, many ARA studies have used textbook data (e.g., Berendes *et al.*, 2018; François and Fairon, 2012; Heilman *et al.*, 2007; Pilán *et al.*, 2016). Another important type of education materials for ARA are leveled articles from news and magazines which were professionally adapted by experts to address learners at different proficiency levels (so called 'graded readers' in the terminology by Vajjala, 2022). The WeeBit corpus (Vajjala and Meurers, 2012) is a prominent example of a readability corpus that was compiled from graded readers. Also the two new readability corpora for German that I present in this thesis are based on graded readers (see Section 5.3.1). A special sub-type of graded readers are 'paired graded readers' (Vajjala, 2022). They feature the same article at different proficiency levels rather than different articles for different proficiency levels. Prominent examples of paired graded readers turned into readability corpora are the Newsela corpus (Xu *et al.*, 2015), or the OneStopEnglish corpus (Vajjala and Lučić, 2018).

Often a single resource does not provide enough leveled reading materials or covers only a small range of reading levels. To address this limitation, it is common practice researchers have combined texts from different resources. For example, the WeeBit corpus combines web materials from the WeeklyReader magazine (web page no longer available) and the BBC-Bitesize website (`http://www.bbc.co.uk/bitesize`) to increase the number of articles and the target populations' age range it covers. Also the ReadingDemands corpus (Vajjala, 2015) combines reading passages from four different textbook publishers to augment the available data. In more extreme cases, leveled corpora can be compiled by combining materials from a resource focusing on a specific target population with materials from a comparable resource focusing on another target population to compose an artificial reading scale. For example, the German Klexikon data set combines articles from an encyclopedia that targets 8–13 years old children (`https://klexikon.zum.de/`) with Wikipedia articles to obtain a binary corpus that can be used for readability assessment, text simplification, and summarization (Aumiller and Gertz, 2022). When combining materials from several data sources it is crucial to minimize any difference across data sources that is not related to the reading level (such as genre or mode): The procedure risks introducing idiosyncrasies from the different data sources that

can partially or fully confound with the differences in reading levels. For ARA models trained on such data, it is particularly important to confirm the generalizability of models in cross-corpus studies because this allows to confirm that the model in fact learned to distinguish the differences between readability levels and not other irrelevant differences between the data sources. That being said, cross-corpus testing is in general important for predictive models to test their generalizability within and across their training domain (see discussion in Section 2.2.3.2). However, comparable corpora for cross-corpus studies are often not available for ARA for languages other than English. In these cases, researchers often use cross-validation to estimate the variability and generalization error on the readability corpus they have (Collins-Thompson, 2014, p. 113). This allows to some extent to test models for overfitting, but it cannot confirm the generalizability of models to other samples from the same population. It also does not identify if models trained on corpora compiled from different data sources learned the intended readability differences. More cross-corpus and cross-domain testing is therefore a central desideratum for ARA research (see also Vajjala, 2022).

While professionally leveled reading materials have many advantages for ARA, they are difficult to procure and share due to legal restrictions and licensing concerns (Collins-Thompson, 2014; Vajjala, 2022). Thus, many researchers have used leveled reading materials that were not rated by experts. One of the most prominent examples of such a corpus is the Wikipedia-Simple Wikipedia corpus. The corpus was a popular resource for text simplification (for an overview, see Siddharthan, 2014) until several studies criticized Simple Wikipedia for insufficiently adapting its language to low literate target audiences (e.g., Štajner *et al.*, 2012; Xu *et al.*, 2015; Yaneva *et al.*, 2016). A more recent example is the previously mentioned Klexikon data set. Vajjala (2022) cautioned against relying on readability levels that were not assigned by experts, pointing out that because of their unknown quality they might align poorly with their intended target audience. However, also publishers of education materials have been shown to insufficiently align the text characteristics of their materials to their intended target audience. Berendes *et al.* (2018) found that German textbook publishers did not systematically develop the linguistic complexity of geography texts across grade levels (5th to 10th grade) and school types (academic secondary school track and vocational secondary school track). Another potential issue with leveled reading materials in general is that publishers and other content providers often themselves use readability formulas or text complexity estimates to inform their readability ratings. This runs the risk of circularity when using these labels to train ARA models based on text characteristics (Collins-Thompson, 2014). The annotation

validity of previously assigned readability labels is therefore always a concern when working with leveled reading materials.

Also model validity is an important concern in ARA research and it is closely linked to the annotation validity of readability corpora: The construct validity of any model that was trained with supervised machine learning methods depends on the construct validity of its reference annotations (see Section 2.2.3.1, pp. 53–54). Some studies used reading times and reading comprehension tests (e.g., Crossley *et al.*, 2014c; Vajjala and Lučić, 2019) as well as eye-tracking (e.g., Gonzalez-Garduno and Søgaard, 2018; Singh *et al.*, 2016; Vajjala *et al.*, 2016) to obtain empirically grounded readability estimates. However, more frequently corpora are annotated with human judgments, either provided by trained expert annotators or by non-expert annotators (see Collins-Thompson, 2014). Non-expert annotations of readability levels have become increasingly common in recent years. They are mostly based on crowd-sourcing experiments (Crossley *et al.*, 2019; De Clercq and Hoste, 2016; De Clercq *et al.*, 2014; Mohammadi and Khasteh, 2020), but also based on user studies pooling multiple non-expert judgments. For example, Naderi *et al.* (2019b) obtained non-expert readability judgments for individual sentences from German L2 readers at different proficiency levels which they averaged into a single opinion score per sentence. More research is still needed on confirming the annotation validity of readability corpora (Vajjala, 2022), especially for languages other than English. Despite increasing interest in these types of annotations for training ARA models, most research continues to use the intended target audience or publisher information as reference labels (Vajjala, 2022; Vajjala *et al.*, 2016). Against this background, Vajjala (2022) identified more large readability corpora with high quality annotations as a central desideratum for ARA research, not only to support training new models and cross-corpus validation, but also to develop best practices in ARA research and to benchmark new models.

Two other factors related to data resources play an important role for ARA: the text unit for which readability is assessed and the reading scale used. ARA has near-exclusively focused on readability assessment for full texts (Collins-Thompson, 2014), but for notable exceptions, see Dell'Orletta *et al.* (2011), Pilán *et al.* (2016) and Vajjala and Meurers (2014) as well as Weiss and Meurers (2022) in this thesis (Section 5.3.4). Even though readability at the text level is important for a variety of applications (see Section 2.3.2), the analysis of smaller units has been repeatedly identified as a desideratum or ARA research (e.g., Pilán *et al.*, 2016; Vajjala and Meurers, 2014). Sentence-level readability assessment can help to identify text passages that require simplification, thus promoting a targeted approach to text simplification.

Furthermore, sentence-level readability models can be applied to short text types such as exercises, social media or chat language, dialogue turns, captions, or questionnaire items. The existing studies showed that the relationship between the overall readability of a text and the readability of its smaller units (paragraphs, sentences) is complex. Difficult texts often contain simple sentences and easy texts can contain difficult sentences (Pilán *et al.*, 2016; Vajjala and Meurers, 2014; Weiss and Meurers, 2022). Similarly, the role a single sentence plays for the overall readability of a text is determined by several factors such as the redundancy of information being encoded by this sentence and the relevance of the sentence for the current reading goal(s). More research is needed to better understand this interplay.

Finally, the reading scales used in readability corpora heavily influence supervised ARA models. There are two central types of reading scales: coarse-grained and fine-grained scales, see Collins-Thompson (2014, p. 102) for a similar distinction. Coarse-grained scales estimate the readability of texts based on broad categories (such as grade levels or CEFR levels but also adult/child). Coarse-grained scales are often intuitively interpretable scales but lack specificity. They are also often specific for a target population (e.g., literate L1 readers, literate L2 readers, or low literate L1 readers). According to Collins-Thompson (2014), grade levels are the most commonly used scale to approximate text comprehensibility. They are the "standard unit of reading difficulty" (Collins-Thompson, 2014, p. 102). This comes from both the use of leveled reading materials which often refer to grade levels (such as textbooks) and the historical focus on readability for L1 contexts (Collins-Thompson, 2014; Sung *et al.*, 2015; Xia *et al.*, 2016). However, grade levels are not suited as scales for L2 readers because L1 readers and L2 readers differ in their abilities (see Section 2.3.1.1). For an adult L2 reader, a text that is suited for a higher grade level (e.g., university-level expository text) may be more comprehensible than a text suited for younger readers (e.g., a novel for children), making a grade scale difficult to interpret. L2 readability assessment often relies on coarse-grained L2 proficiency scales or estimates such as 'beginner', 'intermediate', and 'advanced' or the CEFR scale (e.g., Pilán *et al.*, 2016; Sung *et al.*, 2015; Xia *et al.*, 2016). For low literate readers, readability is typically estimated using binary labels such as '±simplified' (for an exception using a more fine-grained scale, see Weiss *et al.*, 2018). Coarse-grained scales are intuitive for laypeople and relatively straightforward to obtain for researchers and practitioners in need of labeled training data. However, these scales are limited in their adaptability to the skills of an individual reader: By estimating the fit for an average representative of a level, coarse-grained approaches lack the sensitivity to distinguish skill gradients within and between proficiency

levels (Collins-Thompson, 2014, p. 102). For example, an ordinal estimate on the CEFR scale fails to account for texts that fall between two adjacent proficiency levels (e.g., B2 and C1) and does not allow to rank texts falling into one proficiency level (e.g., C1), which is important seeing that learners within the same proficiency level are not necessarily homogeneous in terms of their skills (see Section 2.2.1). In contrast, fine-grained scales are suited to provide a gradient estimate that can be used to distinguish within categories. They are often continuous and allow for a more differentiated assessment of reading skills within a coarse-grained level. For example, a fine-grained scale might estimate on a 5-point Likert scale how readable a text is for an intermediate L2 learner. Likert scales are often used for these kind of more fine-grained readability estimates which are often elicited in psychological or psycho-linguistic studies focusing on experiment-based readability assessment. However, the greater nuance often comes at the expense of interpretability because the scales cannot directly be mapped to the labels used in education research and practice to measure proficiency (Sung *et al.*, 2015, p. 375).

2.3.3.3 Neural and feature-based approaches

Machine learning-based approaches to ARA can be separated into feature-less neural network-based approaches and feature-based machine learning approaches. Currently, both neural and feature-based approaches to ARA report to achieve state-of-the-art performances. For example, Martinc *et al.* (2021) reported an accuracy of 78.7% on the OneStopEnglish corpus (Vajjala and Lučić, 2018) when training and testing different supervised and unsupervised neural models. This is comparable to the performance of the feature-based classification approach by Vajjala and Lučić (2018) which achieves an accuracy of 78.1% and below the accuracy of the feature-based models proposed in Bengoetxea *et al.* (2020, $acc. = 90.1\%$) and Weiss *et al.* (2021, $acc. = 92.1\%$, see Section 5.3.3). On the WeeBit corpus, Mohammadi and Khasteh (2019) reported an accuracy of 91.0% and a RMSE of 0.11 using deep reinforcement learning. Meng *et al.* (2020) achieved an accuracy of 91.7% using a hierarchical self-attention model. This exceeds the accuracy reported for other neural and feature-based approaches (Deutsch *et al.*, 2020; Martinc *et al.*, 2021; Xia *et al.*, 2016) by 5–10%. The results are comparable to the SOTA performance by Vajjala and Meurers (2012) who reported an accuracy of 93.3% an a RMSE of 0.15. However, the neural approaches do not outperform the linguistically broadly informed approach by Vajjala and Meurers (2012). The comparability of feature-based and neural methods in terms of their performance has also been reported for other NLP tasks

(Rigutini and Algherini, 2022).

This shows that neural approaches are not always superior to more traditional feature-based machine learning approaches. It is more accurate to state that neural and feature-based approaches have different advantages and disadvantages. Neural network-based approaches are known to yield high performing predictive models (e.g., Martinc *et al.*, 2021; Meng *et al.*, 2020; Mohammadi and Khasteh, 2019) without requiring the resource intensive process of feature engineering. This is especially important for languages for which the NLP tools needed to extract elaborate linguistic feature sets are not available (Imperial, 2021). However, neural approaches require generally large quantities of data and substantial computational power for training in contrast to feature-based approaches (Bender *et al.*, 2021; Henderson *et al.*, 2020; Rigutini and Algherini, 2022). The choice between neural and feature-based machine learning approaches for ARA needs to consider the available resources. Assuming comparable performance and availability of feature-extraction resources, feature-based models are more resource efficient in terms of energy consumption and training data needed. Especially the latter is often a limiting factor for work on languages other than English or ARA for specific target groups for whom too little data are available, given that there are not enough high quality training corpora for readability assessment (see Section 2.3.3.2).

Feature-based approaches are also generally more interpretable (see also earlier discussion in Section 2.2.3.3), whereas neural approaches currently provide little insights into their decision process. Few contributions to ARA that utilized neural approaches have attempted to linguistically interpret their predictions. A notable exception to this is the work by Madrazo Azpiazu and Pera (2019). They used the attention mechanism in their deep learning model to investigate the linguistic properties of the parts of the texts that receive most attention. Their findings regarding the POS, frequency, and morphological properties of relevant text passages align with previous work on ARA. More research in this direction is needed to make neural approaches to ARA more interpretable. It should be noted that for some use cases, this lack of interpretability is less of an issue (e.g., for user experience and quality control, see Section 2.3.2.3). However, it is a severe limitation in education contexts and for publishers or other content providers who want to use ARA to adapt their materials until they align with their target audience. In this respect, neural approaches parallel with readability formulas, which also do not allow linguistically informed insights for which they have been heavily criticized (Collins-Thompson, 2014, p. 104).

In other ARA contexts, neural approaches have more systematic advantages. For example,

there has been increasing interest in multi-lingual and cross-lingual readability assessment. Many ARA studies train and test their models on multiple languages but do not perform cross-lingual testing (e.g., Imperial, 2021; Martinc *et al.*, 2021; Mohammadi and Khasteh, 2019). Most of these use neural approaches because it is challenging to compute features that can be extracted across languages. Shen *et al.* (2013) are a notable exception to this. They proposed a multi-lingual approach to ARA based on surface length measures and term frequency-inverse document frequency (TF-IDF) and demonstrated the applicability of their approach to Arabic, Dari, English, and Pashto. De Clercq and Hoste (2016) used surface length, lexical, syntactic, and semantic features as well as measures of cohesion for English and Dutch readability assessment and compared their informativeness. Fewer studies work on cross-lingual readability assessment. Research on cross-lingual readability assessment attempts to compensate for the lack of leveled data for certain languages by augmenting it with comparable leveled data from a high-resource language (e.g., Madrazo Azpiazu and Pera, 2020b) or by applying an ARA model trained on one language to another language, treating language differences as a form of domain-adaptation (e.g., Madrazo Azpiazu and Pera, 2019, 2020a; Weiss *et al.*, 2021). Madrazo Azpiazu and Pera (2019) used a multiattentive recurrent neural network trained on English data and demonstrated its applicability across a range of data sets and languages (English, Spanish, French, Italian, Basque, Catalan, and Dutch). Madrazo Azpiazu and Pera (2020b) studied the feasibility of feature-based cross-lingual readability assessment for English, Spanish, Basque, Italian, French, and Catalan. They used surface-based, syntactic, morphological, and semantic complexity features and measures of cohesion to distinguish between simplified and regular encyclopedic texts using Wikipedia and Vikidia (`https://en.vikidia.org`), an online encyclopedia for children (aged 8 to 13 years). Beyond comparing the importance of feature domains across languages, Madrazo Azpiazu and Pera (2020b) demonstrated that they could improve the performance of models for low-resource languages by augmenting the training data with comparable data from other languages. In a follow up study, Madrazo Azpiazu and Pera (2020a) demonstrated that a neural approach utilizing word and sentence level cross-lingual embeddings achieves even better results in cross-lingual transfer.

There has been some work on combining feature-based and neural approaches (Deutsch *et al.*, 2020; Imperial, 2021) but the reported results have been mixed. Imperial (2021) compared use of BERT embeddings, readability features, and a combination of both on several data sets for Filipino and English ARA. They found a clear improvement when combining

BERT embeddings and linguistic features across data sets. In contrast, Deutsch *et al.* (2020) found that augmenting deep learning models using linguistic features only improved model performance on smaller training sets but not when sufficient training data was available. A direct comparison between these hybrid approaches and the SOTA on OneStopEnglish and WeeBit is not possible because both studies report (weighted) f1 scores but not accuracy and RMSE as other ARA approaches on these corpora did. It thus remains unclear how competitive these hybrid approaches are. More research is needed on this combination of approaches. Until then, feature-based and neural approaches to ARA continue to be both relevant for research and practice and methods should be chosen based on the specific application purpose and the availability of resources.

Feature-based approaches have utilized a broad range of textual features that can be connected to the CAF triad and the complexity domains discussed in Section 2.1.2. The choice of features has been shown to be a key factor in model performance (Kate *et al.*, 2010) and studies repeatedly found that the combination of features from multiple linguistic dimensions yielded more robust and accurate models for ARA (e.g., Pilán *et al.*, 2016; Xia *et al.*, 2016), especially for sentence-level ARA (Pilán *et al.*, 2016). Traditionally, readability features have focused on easily observed surface properties of text, such as text, sentence, and word length (see Collins-Thompson, 2014; Feng *et al.*, 2010; Vajjala, 2022). These features were derived from early readability formulas (see Section 2.3.4). Due to the proximity to psychology and psycho-linguistics research on discourse and text comprehension (see Section 2.3.1), vocabulary frequency measures (e.g., Feng *et al.*, 2010; Sung *et al.*, 2015) and cohesion (e.g., Feng *et al.*, 2010; Glöckner *et al.*, 2006; Pitler and Nenkova, 2008; Sung *et al.*, 2015; Todirascu *et al.*, 2013) have also been systematically used in feature-based approaches to ARA. Also features of clausal and phrasal complexity (e.g., Feng *et al.*, 2010; Glöckner *et al.*, 2006; Kate *et al.*, 2010; Pilán *et al.*, 2016; Sung *et al.*, 2015) and lexical complexity (e.g., Feng *et al.*, 2010; Glöckner *et al.*, 2006; Pilán *et al.*, 2016; Sung *et al.*, 2015) have a long tradition in research on ARA. Yet, one of the first explicit references to CAF in computational linguistic work on ARA was made by Vajjala and Meurers (2012). Linking features to the CAF framework and SLA research is important because it helps linking linguistic insights into readability with research on proficiency (Sung *et al.*, 2015). There have also been some early approaches focusing on morphological complexity features in work on languages other than English (e.g., Glöckner *et al.*, 2006; Hancke *et al.*, 2012; Madrazo Azpiazu and Pera, 2020b; Pilán *et al.*, 2016). Most of these studies focused on measures of inflection, derivation, and compound-

ing. Similar to research on ATS, semantic complexity has also played an important role in ARA research including measures of language models, as well as vocabulary ambiguity and abstractness (e.g., Feng *et al.*, 2010; Glöckner *et al.*, 2006; Kate *et al.*, 2010; Pilán *et al.*, 2016; Sung *et al.*, 2015). Measures of human language processing have played only a minor role in ARA, despite their close connection to reading speed and eye-tracking research (Gibson, 2000; Shain *et al.*, 2016) and early calls to include such measures (Bailin and Grafstein, 2001, pp. 294–296). Howcroft and Demberg (2017) used measures of integration costs, surprisal, and embedding depth for sentence-level readability assessment of English. However, they did not test these features in combination with other common readability features. All ARA studies in this thesis used human processing measures for readability assessment (Weiss and Meurers, 2018, 2022; Weiss *et al.*, 2021, Section 5.3). For a more elaborate detailed historical overview of the different types of features used in ARA, see Collins-Thompson (2014).

Some approaches to ARA have also experimented with adding topic modeling and sentiment analysis to their ARA models and to combine these with readers' domain knowledge to bridge the gap between modeling text characteristics and reader characteristics (e.g., Honkela *et al.*, 2012). Similarly, there has been some limited work on informing ARA models by reader characteristics such as proficiency (Collins-Thompson *et al.*, 2011; Tan *et al.*, 2012). As Vajjala (2022) pointed out, though, such features are rare in ARA approaches and more work would be needed on capturing the interplay between text and reader characteristics (see also Bailin and Grafstein, 2001, p. 296). Vajjala (2022, pp. 7–8) also recently advocated for a multi-modal approach to ARA that takes into account not only linguistic text characteristics but also non-linguistic text characteristics such as tables, graphics or a texts' layout. Some studies on ARA for web pages have already utilized non-linguistic information about links, traffic, topic, and references to inform their readability estimates (e.g., Akamatsu *et al.*, 2011; Gyllstrom and Moens, 2010), see also Collins-Thompson (2014, pp. 118–121) for a detailed discussion.

2.3.4 A brief remark on readability formulas

Although the statistical methods discussed in the previous section are the current SOTA in ARA, in practice traditional readability formulas are still widely distributed (Benjamin, 2012; Vajjala, 2022). They continue to be used in numerous applications, for example to screen or filter texts (e.g., Chinkina and Meurers, 2016; Miltsakaki and Troutt, 2008), to evaluate NLG models (e.g., Agrawal and Carpuat, 2019; Marchisio *et al.*, 2019; Stymne *et al.*, 2013), or

to assess the accessibility of web materials (e.g., King *et al.*, 2003; Misra *et al.*, 2013; Paul *et al.*, 2021; Skierkowski *et al.*, 2019). In fact, most studies discussed in Section 2.3.2 used readability formulas to measure readability. Readability formulas focus on easily identifiable surface level text characteristics that are highly correlated with lexical and morpho-syntactic difficulty: word and sentence length (which could be identified without linguistic analysis through graphematic markers such as white spaces and punctuation marks) and curated word lists of either difficult or simple words or frequencies. This makes them easy to use and to fit to new languages. As Collins-Thompson (2014) pointed out, readability formulas were traditionally first designed for English before being adapted for other languages (for an overview of non-English formulas, see Zakaluk and Samuels, 1988). There are numerous overviews documenting the history of readability formulas and cataloging the most influential formulas. DuBay (2004, 2006) provided a comprehensive review of the last 100 years of research on readability formulas. Benjamin (2012) compared readability formulas with early machine learning approaches to ARA in her overview, focusing on their applicability in practice.

Despite their popularity, readability formulas have been heavily criticized for their simplicity and their lack of linguistic awareness. First, they are indifferent to changes in word order, semantic and pragmatic differences, or discourse properties of texts. Any random permutation of a fixed set of sentences will receive the same readability score under a traditional readability formula even though changes in sentence order will heavily impact any text's readability (Bailin and Grafstein, 2001; Benjamin, 2012; Collins-Thompson, 2014). Hence, readability formulas are known to have poor construct validity. Second, their lack of linguistic insight also makes them unsuited as guides for revising and adapting texts: the process of "writing to the formula" has been heavily criticized as ineffective (Benjamin, 2012; Schriver, 2000). This limitation makes them unsuited for educational contexts (Collins-Thompson, 2014; Glöckner *et al.*, 2006). Third, it was suggested that their reliance on surface proxies of text characteristics makes them unsuited for generalizing from traditional texts to non-standard or web data (Collins-Thompson, 2014; Feng *et al.*, 2009). This is especially concerning seeing that they continue to be used to study the accessibility of web materials as discussed above. Fourth, machine learning models typically outperform traditional readability formulas on most data sets (Collins-Thompson, 2014), especially for short texts (Benjamin, 2012). Due to these limitations, ARA research uses readability formulas as baseline models to test the performance of new ARA models (e.g. Pilán *et al.*, 2016). However, outside of ARA research, readability formulas continue to be used in practice because they are well-known, easy to access, and

easy to interpret. Benjamin (2012) evaluated approaches to ARA in her overview not only based on their ecological validity, their fit for a specific target population, and their fit for a specific text type. She also discussed how much training or instruction is needed to use a model and if it is available for immediate use (as opposed to requiring to contact the author of a paper, for example). Readability formulas are easy to access and apply. In contrast machine learning-based approaches are often not accessible to the research community or practitioners in the sense that code and trained models are rarely shared (Vajjala, 2022) and even if they are, they require advanced programming or statistical skills (Benjamin, 2012), because they are rarely integrated in publicly available web platforms. So while readability formulas lack robustness, accuracy, validity, and interpretability, they outperform most SOTA approaches to ARA in terms of accessibility. Despite this mismatch, traditional readability formulas remain the dominant metric of readability outside of ARA research.

Chapter 3

Automating German complexity modeling

3.1 Overview of complexity analysis systems

This chapter introduces the two automatic complexity analysis systems that I used throughout this thesis: the legacy system and the multilingual CTAP system (with a primary focus on its German component). I used the legacy system for linguistic complexity analyses of German throughout this thesis for all but two publications (i.e., Weiss and Meurers, 2022; Weiss *et al.*, 2021). The German component of the CTAP system (i.e., Chen and Meurers, 2016; Weiss *et al.*, 2021) is part of the multilingual analysis extension that I created for CTAP (see Section 3.3) and has a web-based user interface that makes these analyses accessible for a broader audience. I used CTAP in two publications that are part of this thesis (Weiss and Meurers, 2022; Weiss *et al.*, 2021). Both systems use an elaborate NLP pipeline to compute a rich collection of more than 400 features. The feature collection combines complexity measures of absolute and relative complexity (see Section 2.1.1.2, p. 12) in the domains of syntactic, lexical, semantic, morphological, discourse, and human processing complexity (see Section 2.1.2). The systems thus integrate complexity measures stemming from a variety of research areas, including research on SLA complexity, writing quality assessment, text readability, and psycho-linguistic research on human processing (see Chapter 2).

The remainder of this section (3.1) continues with a general non-technical and conceptual overview for the automatic approach used in this thesis. I briefly motivate the need for automatic linguistic modeling and outline the shared conceptual analysis workflow of both systems. The remaining two sections elaborate on the technical details of both systems: Section 3.2 presents the architecture, frameworks, and tools used in the legacy system. Section 3.3 presents the architecture, frameworks, and tools used in the multilingual CTAP system and

also elaborates on how I ensured the comparability of features across languages.

3.1.1 Automating linguistic modeling

The computational linguistic approach presented in this dissertation addresses the call to view and study linguistic complexity as a multi-dimensional construct spanning diverse linguistic domains (Housen *et al.*, 2019; Kuiken *et al.*, 2019; Norris and Ortega, 2009). This allows to investigate trade-off effects between complexity (sub-)domains and to quantify developmental and task variation from different linguistic perspectives (see Section 2.1.3). It contributes directly to overcoming the traditionally reductionist focus of (SLA) complexity research on the syntactic and lexical domain (see Section 2.1.2). This is made possible by automating feature extraction through the use of NLP techniques: Calculating hundreds of features covering several linguistic domains becomes feasible for large quantities of data because an elaborate NLP pipeline and extraction rules fully automate the calculation of complexity measures. While manual expert annotations allow a high degree of control and customization for targeted linguistic constructs (e.g., Bulté and Housen, 2014, p. 48), automatic analyses allow to quickly and efficiently annotate data with numerous linguistic constructs and features at low cost (Crossley and McNamara, 2014; Kyle and Crossley, 2018; Lu, 2010). These annotations are scalable, reproducable, and comparable across studies (Crossley and McNamara, 2014; Lu, 2010; Ströbel *et al.*, 2020, p. 738) while showing good performance on L2 data (Lu, 2010; Ströbel *et al.*, 2020; Weiss and Meurers, 2021). This makes automated approaches a valuable addition to the methodological palette for linguistic research.

The CTAP system makes the broad linguistic complexity modeling approach proposed in this thesis accessible to a broad user base. Similar systems that connect an automated analysis pipeline to a graphical or command-line-based user interface exist primarily for English, such as Coh-Metrix (McNamara *et al.*, 2010a), the Educational Scoring Toolkit (ESCRITO; Zesch and Horbach, 2018), the Lexical Complexity Analyzer (LCA; Lu, 2012), the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010), or the Tool for the Automatic Analysis of Text Cohesion (TAACO; Crossley *et al.*, 2016c). Less work has been dedicated to other languages, but see, for example, Coh-Metrix-Esp (Quispesaravia *et al.*, 2016) for Spanish and Coh-Metrix-Port for Brazilian Portuguese (Scarton and Aluısio, 2010). Recently, there has been increasing interest in providing multilingual systems, but they are still rare due to the technical and conceptual challenges arising from the support of multiple languages. Noteworthy exceptions are MultiAtzerTest for English, Spanish and Basque (Bengoetxea and Gonzalez-Dios, 2021)

and ReaderBench for English, French, Romanian, Dutch, Spanish, Italian, and Latin (Dascalu *et al.*, 2018). For German, however, no comparable systems other than CTAP are available at the time of writing. An early system for the analysis of German was the DeLite system (Brück *et al.*, 2008) but it is not maintained and accessible any more at the time of writing. The Linguistic Analyzer for Text and Item Characteristics (LATIC; Neri and Klückmann, 2021) is currently being developed for English, French, German, and Spanish. However, at the time of writing it calculates fewer and more low level linguistic features than the systems presented here. The focus lies on traditional readability formulas and counts that can be derived from POS tags provided by CoreNLP pipeline (Manning *et al.*, 2014). More elaborate features and a systematic conceptual connection to existing work on ATS, ARA, and (SLA) complexity research is missing. The systems presented in this chapter fill this research gap for German and contribute to work on multilingual analysis systems.

3.1.2 General complexity analysis workflow

Despite their differences in architecture, the legacy system and the multilingual CTAP system conceptually follow the same workflow to extract linguistic complexity measures from raw text data. This is displayed in Figure 3.1. Before we consider the detailed technical implementation of each of these modules in the legacy system (Section 3.2.1 to 3.2.4) and CTAP (Section 3.3.1 to 3.3.4), let us briefly consider the general workflow of both systems for better overview. Both systems start with the **input processing module** which takes an arbitrary

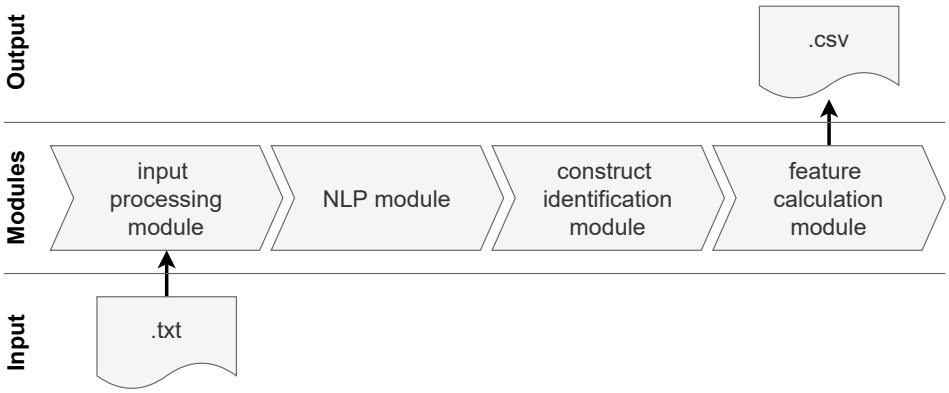


Figure 3.1: *Conceptual analysis workflow of both automatic complexity analysis systems used in this thesis: from plain text to broad linguistic modeling.*

number of plain text input files which the user provides. The module loads these files into

the system for further processing. In the subsequent **NLP module**, input texts are enriched with linguistic annotations (such as sentence boundaries, POS, morphological features, and dependencies). This module combines a series of different NLP components into an elaborate pipeline (for details, see Sections 3.2 and 3.3). These annotations are then fed into the **construct identification module**. There, they are used to identify a pre-defined list of linguistic structures (such as prenominal modifiers or subordinate clauses). A definition of the most important linguistic units used throughout both systems can be found in Appendix A. The final **feature calculation module** calculates the complexity based on the identified constructs. A complete list and short explanation of all complexity features calculated by either system can be found in Appendix B. The complexity analyses for all input files are returned by both systems in a single comma-separated value (CSV) file.

3.2 The legacy system for German complexity modeling

The legacy complexity analysis system is a Java program (compatible with version 8+) that can be accessed via command line. Its most recent version (Weiss and Meurers, 2021) calculates 400 measures of linguistic complexity. This includes measures of absolute complexity and relative complexity (see Section 2.1.1.2, p. 12). The system calculates measures of absolute theoretical linguistic complexity in the domains of syntax, lexicon, semantics, and morphology. It also includes measures of relative complexity in the domains of discourse, human processing, and lexicon. In the remainder of this thesis, I refer to relative lexical complexity measures as 'language use' measures to better distinguish them from absolute lexical complexity measures. The system was originally designed by Hancke *et al.* (2012) but further developed, updated, and extended considerably throughout several iterations (most notably: Galasso, 2014; Hancke, 2013; Weiss, 2015, 2017). In the context of this thesis, I have maintained and updated the system and implemented the flexible input/output capabilities described below.

Figure 3.2 shows the general workflow of the legacy system. It extends Figure 3.1 adding the unique input/output capabilities of the legacy system. The legacy system produces interim serialized output files (*.ser* in Figure 3.2) for the artifacts of the NLP module and the construct identification module (.ser). These are used as input for the subsequent module. This division allows users to start their analysis at different points when they re-run an analysis. Users can start their analysis not only with the input processing module (requiring plain text input) but alternatively also with the construct identification module (requiring the serialized output from

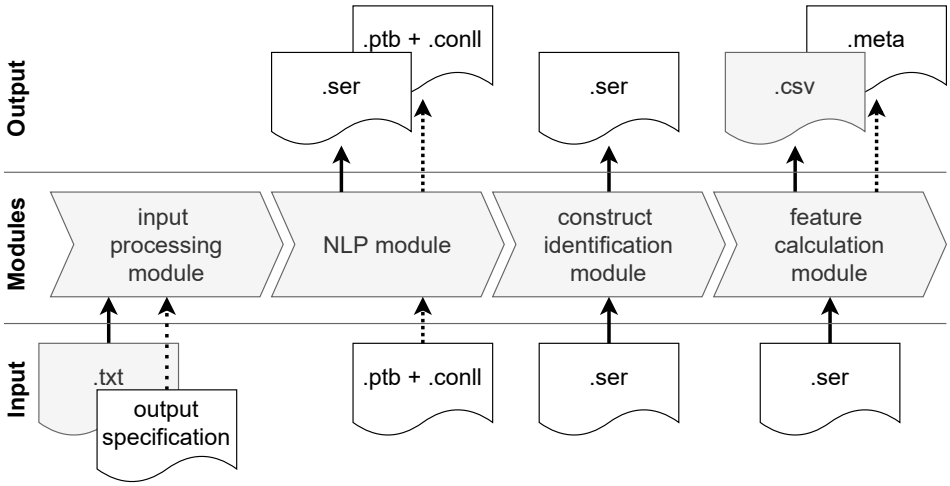


Figure 3.2: *Analysis workflow and input/output capabilities of the legacy system. Additions ot the conceptual workflow presented in Figure 3.1 are printed in white. Optional components are connected with dashed arrows.*

the NLP module as input) or with the feature calculation module (requiring the serialized output from the construct identification module as input). This makes re-running the analysis after altering or adding complexity features more efficient because the NLP module can be skipped. To enhance the interpretability of the linguistic analysis, I added constituency parses—in Penn Treebank (PTB) format—and dependency parses—in Conference on Computational Natural Language Learning (CoNLL) format—as output options for the NLP module. These can be used to inspect the quality of the automatic analysis. I further extended the NLP module to accept linguistic annotations in these formats as alternative input instead of plain text data. When this option is chosen, the module loads the linguistic annotations provided by the user and incorporates them into the output produced for the next module instead of running the NLP pipeline. The input processing module is also omitted in this case. This enables users to use their own annotations for the construct identification module, for example to use gold standard annotations instead of potentially noisy automatic annotations.[1] Users can control these options using command line flags which can be provided as an optional input when starting the analysis. This way, users can define which modules they want to execute and whether or not they wish to create the interim output. By default, all modules are executed and no linguistic annotations are saved except for the serialized files which are always created. Users can also specify whether they want to calculate features on the document level (default), on

[1]For an example of this use case, see Weiss and Meurers (2021) described in Section 5.2.4.

the sentence level, or on the word level (or any combination of those three). This is a second extension of the legacy system that I implemented in the context of this thesis. However, not all features can be meaningfully applied to the sentence level (e.g., global discourse measures) or word level (e.g., TTR).

Optionally, all features can additionally be returned in an attribute-value format (.meta). The optional parse output files (.ptb/.conll) and the attribute-value format file are compatible with the `www.corpus-tools.org` infrastructure for multi-layer linguistic corpora (Druskat *et al.*, 2016). All formats can be converted into the graph-based meta model Salt by using the conversion tool Pepper (Zipser and Romary, 2010). This facilitates, for example, the creation of constituency and dependency annotated complexity corpora in ANNIS (Krause and Zeldes, 2016).

3.2.1 Input processing module

The input processing module iterates over all plain text files (file ending *.txt*) in the input directory provided by the user. Each file is loaded into system memory using an input file stream reader. All texts are read using Universal Transformation Format-8 (UTF-8) character encoding. The input processing module is designed to be easily interchangeable. Users may program their own input processing module suited for customized input formats as long as the output fed into the NLP module is represented in the `AnnotatedDocument` class that is used throughout the system.

3.2.2 NLP module

The NLP module combines several independent NLP components. Sentence segmentation and tokenization are provided by the Apache OpenNLP toolkit (version 1.9.1; `https://opennlp.apache.org/`). The system uses OpenNLP's default model which was trained on the Leipzig corpus (Goldhahn *et al.*, 2012). For compound splitting, the module uses the jWordSplitter (version 3.4; `http://www.danielnaber.de/jwordsplitter/`) which is dictionary-based. POS tagging as well as lemmatization, morphological analysis, and dependency parsing are all provided by the Mate tools (version 3.6; Bohnet and Nivre, 2012). The system uses the Mate tools' default model which was trained on the Tiger treebank for dependencies (Brants *et al.*, 2002) without ellipses, see description in Seeker and Kuhn (2012). The legacy system employs two more parsers: The Stanford CoreNLP pipeline (version 3.9.2; Chen and Manning, 2014)

produces constituency parses for the system. For this, it uses the default model that was trained on the Negra corpus (Brants *et al.*, 1999). The Berkley parser (version 1.7; Petrov and Klein, 2007) produces parses of topological field structures. For this, the system uses the model trained by Ziai (2018) on the TüBa-D/Z treebank (Telljohann *et al.*, 2004). Both parsers use the POS annotations produced by the Mate tools to foster consistency across NLP tools.

3.2.3 Construct identification module

The construct identification module is based on extraction rules and external linguistic resources. The calculation of syntactic complexity measures relies heavily on Tregex (Levy and Andrew, 2006), a tool that applies regular expressions to tree structures. Semantic information is obtained through the GermaNet word net for German (version 11; Hamp and Feldweg, 1997). The module also retrieves single- and multi-word connectives from precompiled lists based on Breindl *et al.* (2014) and Eisenberg *et al.* (2009). These list connectives by type (e.g., temporal or causal) which allows a more fine-grained analysis of how cohesive devices are used. For the calculation of word frequencies, the system utilizes several frequency data bases for German. To capture language use in newspapers and written academic language, it uses the dlexDB data base (Heister *et al.*, 2011). For general spoken and written language use, the system extracts frequencies from Subtlex-DE and Google Books 2000 (Brysbaert *et al.*, 2011). Finally, the system approximates children's language use through frequency and age of active use measures that I extracted from the KCT corpus (Lavalley *et al.*, 2015) in Weiss (2017).

3.2.4 Feature calculation module

The feature calculation module retrieves the linguistic counts that are required to calculate features based on the predefined formulas and algorithms. All results are saved domain-wise in form of hash maps. After feature calculation, the module iterates over these hash maps and combines them into a wide-format table. Depending on the output file name specified by the user, the table is tab-separated (*.tsv*) or comma-separated (*.csv*).

Most features are ratios that normalize the occurrence of the target construction (e.g., the count `number of complex NPs`) by a suitable unit (e.g., sentence, t-unit, clause, NP). Norris and Ortega (2009, p. 560) pointed out that the choice of denominator should be adequate for the data that is being analyzed. To leave the choice of the denominator up to the researchers

using the system and their research questions, the legacy system calculates multiple normalizations of many counts, such as `number of complex NPs per sentence`, `number of complex NPs per t-unit`, `number of complex NPs per clause`, and `number of complex NPs per NP`. The latter is a sub-type of ratio features because it calculates the percentage of complex NPs. The system calculates several percentage features across linguistic domains. The system includes several percentage features for complementary linguistic constructs (dependent and independent clauses or different noun cases). Examples of percentage features are `number of independent/dependent clauses per clause`. The module also calculates several length features. They focus on the length of linguistic constructs (e.g., words, phrases, clauses, t-units, sentences) in terms of different units (characters, syllables, words), again leading to multiple versions of a specific length feature (e.g., `word length in syllables` and `word length in characters`). Again researchers can choose which length feature(s) to use based on their research questions and data. These three feature types (ratio, percentage, length) make up the majority of features. Other features include elaborate formulas or calculation algorithms—such as for MTLD (McCarthy and Jarvis, 2010), PID (Brown *et al.*, 2008), and the DLT measures proposed by Weiss (2017)—or measures calculating the coverage of variants for a specific construct—such as the `coverage of noun modifiers`. Frequency features are measured as raw frequencies and frequencies per million words as well as in terms of frequency bands. Frequency bands are calculated on a Zipf scale (log10) which is partitioned into integer ranges (1–1.$\bar{9}$, 2–2.$\bar{9}$, ...). Frequency band features allow to identify the number of low-frequency and high-frequency words (Brysbaert *et al.*, 2018, p. 46).

3.3 A multilingual common text analysis platform

The multilingual Common Text Analysis Platform (CTAP) is a fully web-based analysis platform for linguistic complexity analyses. CTAP is freely available for immediate use at `www.ctapweb.com`. The currently deployed version of CTAP (Weiss *et al.*, 2021) calculates a total of 1,049 measures of linguistic complexity from the domains of syntax, lexicon, morphology, discourse, human processing, and language use for the languages English (EN; $N = 889$), German (DE; $N = 543$), French (FR; $N = 368$), Spanish (ES; $N = 387$), and Dutch (NL;

$N = 212$), and Portuguese (PT; $N = 1$).[2, 3, 4]

The CTAP system was originally designed by Chen and Meurers (2016) as a monolingual analysis platform for English. It was fully written in Java (compatible with version 8+) and uses Google Web Toolkit (GWT) (version 2.7) for the front-end and Unstructured Information Management Application (UIMA) in the back-end. Front-end and back-end use a shared PostgreSQL database. As part of this thesis, I have extended CTAP to be a multilingual analysis platform to be able to integrate the German features from the legacy system into the existing tool for English. In this context, I have also extended the existing feature sets for English from Chen (2018) and German (see Section 3.2) with the features from the respective other system when possible. To test the generalizability of the multilingual system, I have also started the integration of Dutch. The other languages were added later on by other researchers whom I consulted and are only listed here for documentary purposes. The technical details of the general CTAP architecture has been described in detail in Chen and Meurers (2016) and Chen (2018). In the following, I will only discuss it as far as it is relevant to understand the general workflow or the changes that I made to the system to support multilingual complexity analyses.

Figure 3.3 shows the general CTAP workflow. It extends Figure 3.1 adding the unique input/output capabilities of the CTAP system.[5] Unlike the legacy system, the CTAP modules are activated demand-driven and not executed in a fixed sequence. This means that if a user chooses to only analyze lexical complexity measures, the NLP components required for other features (such as parsers) are not used. To enable this, the input processing module in CTAP is divided into three asynchronously operating components: the corpus manager, the feature selector, and the analysis generator. These are relevant for the input processing module discussed below.

Before elaborating on the different modules, let us briefly review the central UIMA terminology that is needed for the remainder of this section. UIMA was designed to facilitate the analysis of unstructured data. It represents individual analysis units (e.g., text documents) as UIMA Common Analysis Structure (CAS) objects. All processing and retrieval processes that

[2] Even though the NLP pipeline would support the calculation of more features, currently only one feature is publicly available for Portuguese, namely `number of sentences`.

[3] All features are listed and defined within the CTAP platform at `www.ctapweb.com`. A complete list and description of all features for German can also be found in Appendix B.

[4] A version for Italian exists (Okinina *et al.*, 2020) but it has not yet been integrated into `www.ctapweb.com`

[5] The system architecture of CTAP does not explicitly implement these modules. However, the individual implemented components in CTAP can be linked to the conceptual workflow as outlined below.

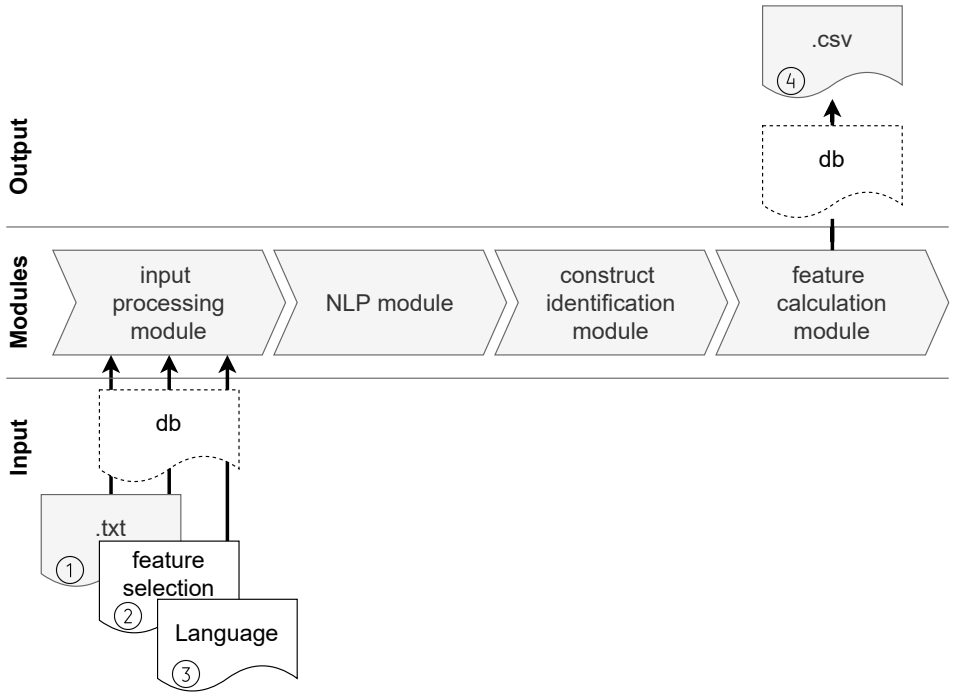


Figure 3.3: *Analysis workflow and input/output capabilities of the CTAP system. Additions to the conceptual workflow presented in Figure 3.1 are printed in white. Dashed boxed indicate that input/output is not directly accessible to the user.*

need to be executed on a text are performed on its CAS representation. CAS objects can be accessed and enriched with additional information using UIMA Analysis Engines (AEs). AEs can add span annotations to the CAS. AEs that interact with the CAS after other AEs have been executed, can build on previous annotations. To enable this step-wise annotation process, AEs utilize so called UIMA Feature Structures to define the types of annotations that they perform and the values these annotations can take. Besides the actual Java code used to perform the task of an AE (e.g., annotating sentence boundaries or named entities), AEs require Extensible Markup Language (XML)-based descriptor files that specify their input/output capabilities and dependencies, links them to a specific UIMA Feature Structure, provides pointers to external resources, and specifies other configuration options.

3.3.1 Input processing module

The corpus manager allows users to upload their plain text data into the CTAP database—(1) in Figure 3.3 and Figure 3.4. Similarly, the feature selector allows users to compile feature

sets from the existing pool of all available features—(2) in Figure 3.3 and Figure 3.4. Data sets and feature sets are not shared between users and are only accessible to their owners. Once they are stored in the database, data sets and feature sets can be used by their owners to assemble any number of analyses in the analysis generator. There, users can create an analysis by combining a data set and a feature set that they own with an analysis language—(3) in Figure 3.3 and Figure 3.4. The user interface prompt for the specification of the analysis demands is shown in Figure 3.4. This specification is used to identify which components of

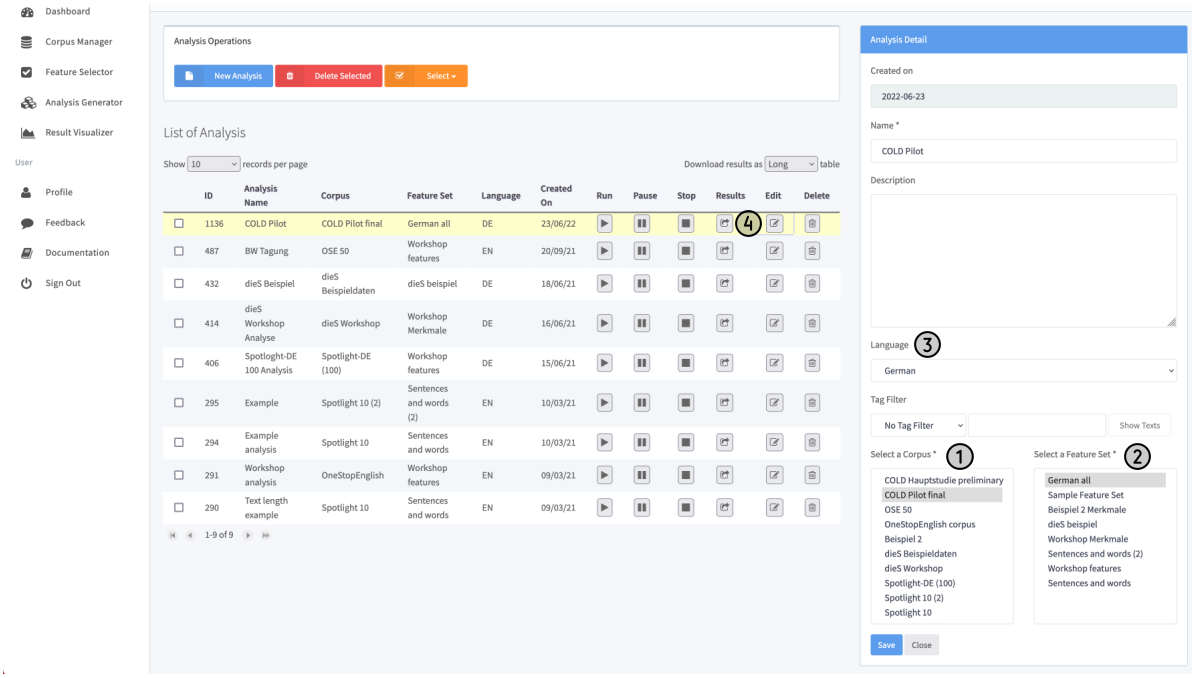


Figure 3.4: *CTAP analysis generator view: central access point for CTAP's input/output capabilities after uploading users' corpus data and defining user-specific feature sets. Numbers 1 to 4 are aligned with those in Figure 3.3.*

the subsequent analysis modules are needed.

3.3.2 NLP module

Overall twelve AEs belong to the NLP module (#0 to #11; henceforth: annotators).[6] They are listed in Table 3.1. I discuss in the following the tools used for all six languages currently represented in CTAP for documentation purposes and to illustrate the flexibility of

[6]For the remainder of this chapter, I use the term 'annotator' exclusively to refer to AEs not to human annotators.

Table 3.1: *Overview of annotators in the CTAP NLP module and the NLP tools and models that they use for English, German, French, Spanish, Dutch, and Portuguese (C = CoreNLP pipeline, O = OpenNLP pipeline, S = stanza pipeline, M = Mate tools, SB = Snowball stemmer, N = do nothing dummy annotator, ∅ = no NLP resources needed, n.a. = not available).*

Annotator	#	Requires	EN	DE	FR	ES	NL	PT
Full pipeline	0	none	N	N	N	N	N	S
Sentences	1	0	C	C	C	C	O	N
Tokens	2	0, 1	C	O	C	C	O	N
Types	3	0–2	∅	∅	∅	∅	∅	∅
Letters	4	0–2	∅	∅	∅	∅	∅	∅
Syllables	5	0–2	∅	∅	∅	∅	∅	∅
POS	6	0–2	C	C	C	C	O	N
Lemmas	7	0–2	M	M	M	M	*n.a.*	N
Stems	8	0–2, 7	SB	SB	SB	SB	SB	*n.a.*
Morphological tags	9	0–2, 7	*n.a.*	M	M	M	*n.a.*	N
Constituency trees	10	0–2, 6	C	C	C	C	*n.a.*	N
Dependencies	11	0–2, 6, 7	C	C	C	C	*n.a.*	N

the multi-lingual architecture that I introduced with the integration of German into CTAP. Eleven annotators are dedicated to the annotation of individual linguistic units (#1 to #11 in Table 3.1).[7] The availability of these annotators for the different languages are listed in Table 3.1. The sentence annotator (#1) uses tools from the Stanford CoreNLP pipeline (version 4.2.0; Chen and Manning, 2014, EN, DE, ES, FR) and Apache's OpenNLP toolkit (version 1.9.1; `https://opennlp.apache.org/`; NL). The token annotator (#2) requires the sentence annotator as a pre-processing step and uses tools from CoreNLP (EN, ES, FR) and OpenNLP (DE, NL). The annotators for types (#3), letters (#4), and syllables (#5) each require the sentence and token annotator as pre-processing steps and do not require additional external NLP tools to identify the respective linguistic units. Unique tokens are identified using string comparisons. Letters and syllables are identified using regular expressions with only the latter being language dependent (assuming languages written in a Latin alphabet). The POS annotator (#6) requires the sentence and token annotator as pre-processing steps and uses tools from CoreNLP (EN, DE, ES, FR) and OpenNLP (NL). The lemma annotator (#7) requires the sentence and token annotator as pre-processing steps and uses the Mate tools (version 3.61; Bohnet and Nivre, 2012). The stem annotator (#8) requires the sentence and token annota-

[7] I will discuss the full pipeline annotator (#0) later in this section.

tor as pre-processing steps as well as the lemma annotator. It uses the OpenNLP Snowball stemmer (EN, DE, ES, FR, NL) on lemmas instead of tokens to improve stemming accuracy. The morphological tag annotator (#9) requires the sentence, token, and lemma annotator as pre-processing steps. It uses the Mate tools. This annotator is currently not available for English and Dutch. The constituency tree annotator (#10) requires the sentence, token, and POS annotator as pre-processing steps and uses tools from CoreNLP (EN, DE, ES, FR). This annotator is currently not available for Dutch. The dependency annotator (#11) requires the sentence, token, and POS annotator as pre-processing steps. It uses the CoreNLP dependency parser (EN, DE, ES, FR). This annotator is currently not available for Dutch. All tools use the respective default models for the different languages.

The division into separate annotators for different linguistic units corresponds to the original separation of annotators in the monolingual English CTAP system by Chen and Meurers (2016) which contained a total of eight annotators. In the context of this thesis, I added the annotators for stems (#8), morphological tags (#9), and dependency parses (#11). Chen and Meurers (2016) chose this architecture because it allows to load only those components and models that are actually required for any given analysis instead of running the full pipeline. In the new multilingual context that I have established with this thesis, this strict division also promotes the flexible exchange of individual NLP components for different languages. This becomes necessary if an NLP pipeline does not support certain annotations for a language. For example, at version 4.2.0, CoreNLP did not offer German-specific models for tokenization and dependency parsing and does not include any morphological analysis component for any language. Tokenization for German is supported by using the English model as default which shows an inferior performance compared to other tokenizers trained specifically for German. To circumvent this issue, the NLP module currently uses the OpenNLP tokenizer and the Mate tools for morphological analyses. A disadvantage of this architecture is the increased programming effort and the lower efficiency for cases where the entire pipeline is needed and a single pipeline can perform all required annotations. For this special case, I have developed another annotator that allows to use a complete NLP pipeline to perform all annotations, the full pipeline annotator (#0). It is currently only used for Portuguese which uses the stanza pipeline (Qi *et al.*, 2020) for sentence segmentation, tokenization, POS tagging, lemmatization, morphological tagging, constituency parsing, and dependency parsing. Also the CoreNLP pipeline has been implemented for the full pipeline annotator but is currently not used.

Adding new languages and the corresponding NLP tools, models, and other linguistic resources to support them to an annotator is prone to introduce code redundancies that hinder the readability and maintenance of annotators. To avoid this, I designed a generalizable class structure that extends the given structure of UIMA AEs. UIMA AEs extend the JCas annotator implementation base (`JCasAnnotator_ImplBase`) from which they inherit several methods. CTAP overrides three of these methods: `initialize()` to load and initialize all required resources and instance variables, `process()` for the actual text-wise annotation, and `destroy()` to free memory after completing the analysis.[8] Additional to these three methods, each annotator in CTAP contains $2+n$ private classes with n being the number of NLP tools used for the annotation: the annotation interface, a dummy annotator that allows to skip the present analysis,[9] and the classes implementing the interface. These classes serve as wrappers for NLP tools and standardize the methods and input/output capabilities across NLP tools to match the interface. The `initialize()` method selects the wrapper that should be initialized based on the language parameter that users selected for the analysis (3). Any language-specific resources that are required are also loaded at this point. The paths to the language-specific resources are stored in the UIMA XML descriptor of the AE. They are retrieved using identifiers that match the provided language codes. This way, no extra code is required to load the correct model because everything is handled through parameter variables.

Let us consider the token annotator (`TokenAnnotator`) as a concrete example: I implemented the `CTAPTokenizer` interface which specifies the abstract method `tokenize()` that takes an input sentence `String` and returns an array of OpenNLP `Span` objects.[10] I then implemented two wrappers implementing this interface: one for the Stanford CoreNLP tokenizer and one for the OpenNLP tokenizer. The wrappers contain all tool-specific code including the conversion from the original output of the CoreNLP tokenizer (an array list of `CoreLabel` objects) to a `Span` array. `TokenAnnotator` has an instance variable of type `CTAPTokenizer`, i.e., the interface. In `initialize()`, the `CTAPTokenizer` variable is initialized as the language-appropriate implementation of this interface through a switch statement again based on the user input language. The language-specific model is loaded at this step, too. The path to the languaguage-specific model is specified in the of type `CTAPTokenizer` XML descriptor. This

[8] The overridden `destroy()` methods only add minor logging capabilities to the inherited `destroy()` method.

[9] This is necessary for two reasons. First, it allows to skip annotators #1 to #11 if the full pipeline annotator is used and vice versa. Second, it introduces flexibility if NLP tools for different languages have different dependencies.

[10] I chose `Span` arrays as output for the tokenizer interface because OpenNLP supports tokenization models for more languages than CoreNLP.

set-up allows to keep the code within `process()` fully language independent. The method only references the methods specified in the interface. It iterates through the sentence annotations that were stored in the UIMA CAS representation of the analyzed text. For each sentence, it calls the `tokenize()` method that is implemented by all wrapper classes, and saves the returned tokens within a `Span` array to the CAS.

3.3.3 Construct identification module

The construct identification module also makes use of UIMA AEs. However, it annotates the CAS with counts rather than linguistic structures. The AEs in this module are executed after the annotators, so that they can use the annotations made in the NLP module as preprocessing steps and access the annotations through the CAS before adding their own annotations in form of construct counts. The `initialize()` method loads all language-specific resources that are still needed. For example, CTAP uses several frequency databases: Subtlex for US English (`https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus/overview.htm`; Brysbaert and New, 2009), German (`http://crr.ugent.be/archives/534`; Brysbaert *et al.*, 2011), and Spanish (`http://crr.ugent.be/archives/679`; Alonso *et al.*, 2011; Cuetos *et al.*, 2012b), the Lexique film frequency database (`http://www.lexique.org/`; New *et al.*, 2004), as well as OpenSubtitles for German, English, French, Spanish, and Dutch.[11] It additionally contains all databases for frequencies, age of acquisition, age of active use, and concreteness that are discussed for German in Section 3.2.3 (p. 93) and for English in Chen (2018). Other language-specific resources used in this step are Tregex patterns (Levy and Andrew, 2006) for constituency trees or the lists of connectives used for German in the legacy system (see Section 3.2).

The `initialize()` method also initializes a language-specific extension of the abstract classes `WordCategories` and `DependencyLabelCategories`, if POS tags or dependency labels are required to identify a linguistic construct. Following a similar logic as the previously described wrapper classes for NLP tools, these classes for POS and dependency labels serve as wrappers for language-specific tag sets. This allows to use the same methods and key words in the `process()` and the XML descriptors for commonly needed word categories (e.g., lexical words, function words, verbs, pronouns, punctuation, etc.) and dependency labels (e.g., root or argument). Besides language-specific extensions, CTAP contains extra classes for POS and

[11] `https://github.com/hermitdave/FrequencyWords/`

dependency tags from the Universal Dependency framework.[12]

All extraction rules that are further used to identify the predefined target constructs are designed to be maximally comparable across languages. This is discussed in more detail in Weiss *et al.* (2021).

3.3.4 Feature calculation module

The feature identification module also makes use of UIMA AEs. As the construct identification module, it annotates the CAS with numeric values rather than linguistic structures. The design of AEs in this module is virtually identical to the AEs in the construct identification module. In fact, within CTAP, no structural difference is made between both modules except that feature AEs require not only AEs from the NLP module as pre-processing steps but additionally build on AEs from the construct identification module. The types of features that are calculated in CTAP are comparable to those calculated in the legacy system (see Section 3.2.4). CTAP calculates ratios, percentages, length features, and coverage features, as well as features based on more complex formulas and algorithms. An important difference to the feature calculation module that is used in the legacy system is that the output of this module is not directly returned as a data table. Instead, CTAP saves all features and raw counts that were requested by the user in the PostgreSQL database. After an analysis has been completed, users can download their results as a CSV file in wide or long format through the web front-end. They can also use the result visualizer to obtain first visualizations of their data if they provided relevant meta information when uploading texts using the corpus manager. For more details on the result visualizer, please see Chen and Meurers (2016) and Chen (2018).

[12] `https://universaldependencies.org/`

Chapter 4

Systematic literature surveys

4.1 Motivation and shared design principles

This chapter reports two systematic, descriptive literature surveys: firstly, one on APA and ATS for German (Section 4.2) and secondly, one on Automatic Readability Assessment (ARA) for German (Section 4.3). In the remainder of this chapter, I use the term Automatic Language Performance Scoring (ALPS) to jointly refer to work on APA and ATS. Assessing the quality of writing covers a broader range of sub-tasks than ARA and the general notion of 'language performance' allows us to also take into account work on writing quality assessment that is at the periphery of APA and ATS research. Examples for such work are the evaluation of the appropriateness of professional communication (e.g., Ludwig *et al.*, 2021) or of age-appropriate social media communication (e.g., Frey, 2020a).

Both topics attract increasing attention from various research disciplines, especially (but not exclusively) in education-related research (for details, see Sections 2.2 and 2.3). The ever-growing pool of interdisciplinary stakeholders in particular is making it increasingly difficult to maintain an overview of the current research landscape. Our two systematic literature reviews aim to address this issue. These are the first literature surveys specifically dedicated to ARA and ALPS for German. To the best of our knowledge, they are also the first *systematic* surveys on either topic for any other language. Unlike traditional narrative reviews that rely on the literature knowledge of their authors, systematic literature reviews follow predefined criteria for the retrieval and inclusion of literature (Xiao and Watson, 2017). A structured survey aims to be as comprehensive as possible within its predefined scope, even if it is unlikely to elicit the full population of relevant publications. This principled way of eliciting literature results in a more comprehensive empirical basis for the literature review. Structured

surveys are less influenced by potential biases of investigators towards certain methods, publication venues, or research groups and result in a more representative sample of the available literature. This strengthens the conclusions that can be drawn from them.

In this chapter, we provide two systematic literature reviews to analyze the current state of research on ALPS/ARA and the diffusion of computational linguistic methods in an interdisciplinary context. We used the PRISMA statement (consisting of a checklist and a flow diagram) as an orientation to ensure the transparency and completeness of our surveys (Page *et al.*, 2021). A direct application of the PRISMA checklist for reporting standards was not possible because the checklist was designed for surveys of intervention studies (Page *et al.*, 2021, p. 1) and is only partially applicable to other types of surveys such as the two presented here. Figures 4.1 and 4.10 in this chapter contain the respective PRISMA flow diagrams for the documentation of the identification, screening, and inclusion of literature for both surveys.

In the present surveys, we focused on the following research questions:

a) Which research disciplines use automatic methods to assess readability / language performance? Which methods do they use?

b) Which types of language and target groups are the approaches trained on and used for?

c) Which machine learning methods and features are used to predict language performance/readability?

d) What is the current SOTA performance for ALPS and ARA models?

e) How available and accessible are SOTA ALPS and ARA approaches?

Benjamin (2012) set out to answer a similar set of questions for her review of ARA for English. However, our surveys put a stronger emphasis on methodological aspects of current work on ALPS/ARA.[1] We answered our research questions with the goal to better understand where research on ALPS and ARA stands in terms of methods and data used and the extent to which it has arrived in (interdisciplinary) practice. This serves primarily to inform researchers about which directions new work on ALPS/ARA for German still needs to explore and what can be done to enhance the impact of this work in practice. In the following, we will first present the shared study set-up for both surveys (see below), before showing and discussing the results for the ALPS survey (Section 4.2) and the results and discussion of the ARA survey (Section 4.3).

[1] Sections 2.2 and 2.3 discuss the conceptual and methodological background that has influenced the perspective of these surveys in more detail.

Literature identification We chose Google Scholar (`www.scholar.google.com`) as literature data base. Unlike other data bases (such as Semantic Scholar), it indexes work across disciplines from a variety of research contexts and includes pre-prints, gray literature, white papers, and graduate or under-graduate theses (Xiao and Watson, 2017, p. 103). It also includes records from a uniquely broad range of research disciplines. This is ideal for the present surveys, because we wanted to reduce the sampling bias from querying research discipline-specific data bases (such as the anthology of the Association for Computational Linguistics). We could not obtain a balanced sample of discipline-specific data bases because uncovering which research disciplines are involved in ARA/ALPS was part of our research questions. Using only one literature data base comes with the trade-off that we cannot find work that is not indexed by Google Scholar. However, we considered this to be an acceptable limitation given the size of the Google Scholar data base. Also, even though a near complete list of the relevant literature would be ideal, a representative sample suffices to answer our research questions.

We wrote a python script to crawl Google Scholar automatically for a broad range of search terms (see details in Sections 4.2 and 4.3).[2] The script extracts for each query result the title, list of authors, text snippet, and links to the paper from Google Scholar and saves these information in JSON format. To facilitate manual screening, the results can also be exported to CSV format. To focus on the development of both fields in the time frame in which computational approaches to performance and readability assessment were available (cf. Sections 2.3 and 2.2), we queried for literature published in the past 20 years (2002–2022). We retrieved literature on ARA on February 16th, 2022 and literature on ALPS on March 10th, 2022. Later work has not been considered in the surveys. From these results, we included the first 10 pages (i.e. the first 200 hits) sorted by relevance in our surveys. We found that the fit between retrieved manuscripts and query terms considerably decreased at later pages while piloting the study design.

Inclusion criteria In both surveys, we only included quantitative empirical studies using ALPS/ARA for German. Qualitative studies and surveys or reviews were excluded. Multilingual approaches were only included, if they were evaluated on German language data. Studies did not need to focus on ARA to be included, as long as they used automatically calculated readability scores that were at least partially inferred based on language properties. We excluded studies that only used manually annotated language properties as features to predict or

[2]The script is available online at `https://github.com/zweiss/crawl-scholar`.

study language performance or readability scores because the focus of both surveys lies on characterizing research on *automatic* approaches to both topics. We focused on analyses of written language in our choice of search terms, but analyses of spoken language were not excluded if they were retrieved. For the ARA survey, only studies that assessed readability with a holistic score at or above the sentence level were included. This excludes work on complex word identification. This restriction did not apply to the ALPS survey because we did not want to exclude SLA research on characterizing developmental processes and performance differences if it used automatically calculated measures. For the ALPS survey, we focused solely on research assessing language performance. Work focusing on the factual correctness of answers was not included. All types of full text publications were considered as candidates in both surveys. Beyond peer-reviewed papers, this also included pre-prints, BA theses, MA thesis, and PhD thesis. We restricted the surveys to papers written in English or German.

Literature screening and included literature Following the automatic identification of candidate literature, we manually screened the results. We removed duplicates and papers for which the full text was not accessible or not written in English or German. The remaining candidates were tested for their adherence to the previously discussed inclusion criteria using a three-step procedure. First, we removed all manuscripts whose title clearly demonstrated that they violated one or more inclusion criteria. This was for example the case for studies where the title explicitly stated a target language other than German.[3] Second, for the remaining candidates, we read the abstract to identify if they violated the inclusion criteria. Third, we read to full text of all remaining papers to determine their suitability for the survey. If a paper fit all inclusion criteria, it was encoded along several categories using a standardized annotation scheme. The categories are:

Research disciplines title, publication venue, publication year, research domain, publication type, primary research goals,[4]

Data and labels $\pm$ cross-lingual approach, language(s), target group, production/reading purpose, text type, corpus name/scale/mode of annotation (separately for train and test data), assessment level (e.g., sentence, paragraph, document)

Methods and metrics general machine learning approach (end-to-end neural networks vs.

[3]For example, “Simple or complex? Assessing the readability of Basque texts”.

[4]For example, to train and test a new readability classifier or to evaluate the readability of certain web materials.

feature-based), number of features (if available), numeric document representation (e.g., name of word embedding or types of complexity features used), type of statistical method (e.g., regression, classification, clustering), name of statistical method, evaluation metric(s), train-test set-up, performance of best model, validation method(s)

Availability ± shares model or code, ± model available for immediate use, ± no prior user training needed to use model

The literature screening and encoding of the final candidates was conducted by one trained annotator who consulted me as the primary investigator in unclear cases. I controlled all annotations at the end of the encoding process. For the few cases of disagreement, the final encoding decision was made by me. The result table for each study additionally contained columns for the unique study ID, annotator, comments, and open questions. The result tables can be found in the online supplementary material to this thesis at `https://osf.io/5vb2x/?view_only=6d1bb8ccfe3f458c946ff4fd6ef5206b`. All analyses based on these annotations were conducted using the statistical computing programming language R (R Core Team, 2022) and the `tidyverse` package (Wickham *et al.*, 2019).

4.2 Automatic proficiency assessment for German: a structured survey of research from 2002 to 2022

The structured ALPS survey followed the general study design outlined in the previous section. Table 4.1 shows the search terms we used to identify relevant literature in Google Scholar. These terms were identified as the central terms for ALPS with a focus on ATS and APA during research for the narrative literature review reported in Section 2.2. Figure 4.1 summarizes the individual steps of the literature screening process. The literature search with these terms yielded a total of 3.734 candidate records.[5] Of these, 285 were removed prior to screening because they were duplicate records and 95 papers could not be retrieved. We screened the remaining 3,354 papers initially based on their titles and abstracts. This way we identified 11 papers written in a language other than English or German. A total of 2,434 were clear content mismatches.[6] We also retrieved several surveys which we excluded following our exclusion

[5] The raw candidate records are available in JSON format in the online supplementary material (`https://osf.io/5vb2x/?view_only=6d1bb8ccfe3f458c946ff4fd6ef5206b`).

[6] An example for a clear content mismatch is Sheldon's (2020) paper titled *'We cannot abandon the two worlds, we have to be in both': Chilean scholars' views on publishing in English and Spanish*.

Table 4.1: *Search terms and patterns used for the structured ALPS literature survey for German (2002-2022, search terms are comma-separated). All search terms are based on the literature commonly used for ALPS which was identified in the context of preparing the background chapter on ALPS (Section 2.2).*

Language	Search pattern	Search terms
English	"`search term`" AND German	*writing competency*, *writing complexity*, *writing proficiency*, *writing quality*, *writing evaluation*, *language competency*, *language proficiency*, *essay grading*, *essay quality*, *essay rating*, *complexity index*, *text assessment*, *text complexity*, *text difficulty*, *text quality*
German	"`search term`"	*Schreib-Kompetenz* (engl. "writing competency/proficiency"), *Schreib-Komplexität* (engl. "writing complexity"), *Schreib-Qualität* (engl. "writing quality"), *Sprach-Kompetenz* (engl. "language competency/proficiency"), *Essay-Rating* (engl. "essay grading/rating"), *Essay-Qualität* (engl. "essay quality"), *Text-Komplexität* (engl. "text complexity"), *Text-Qualität* (engl. "text quality")

criteria ($n = 29$). This left 1,004 papers which entered the full paper screening. Most studies focused on languages other than German ($n = 739$). Overall 120 papers on German ALPS were excluded because they did not report quantitative empirical studies on language performance that were measured with automatically calculated text characteristics. Of the remaining 21 papers, two papers each included not one but two independent studies that were relevant for our survey (Frey, 2020b; Weiss, 2017). We considered studies independent if they addressed different research questions and used different reference labels and data sets. We treated each study as separate paper for the purposes of the survey (Frey, 2020b; Weiss, 2017a,b; Frey, 2020a). This resulted in 23 studies which were included in the present survey.

The low number of papers suited for our survey is somewhat surprising. However, we believe this to indicate a genuine lack of research on ALPS for German within our specified inclusion criteria for the following reasons: Based on our prior non-structured literature survey, the list of search terms comprehensively covers the central terminology used in work on APA and ATS. The terms seem to have been sufficiently precise since 26.4% of records screened were related to ALPS. The stopping criterion at 200 records does not seem to have been too early either. A post-hoc inspection revealed that records ranked at positions 150–200 qualified notably less often for the survey than records with a higher ranking. Also, we

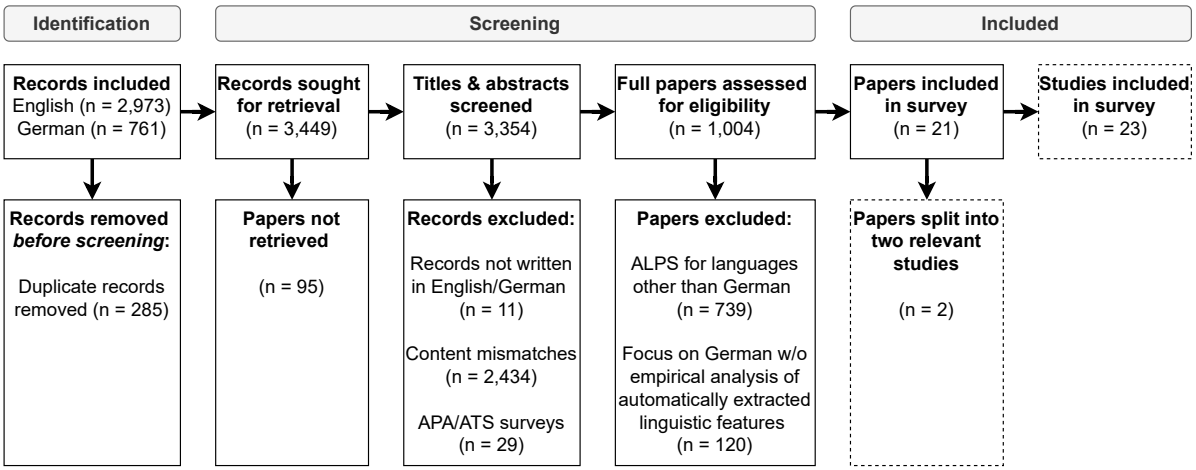


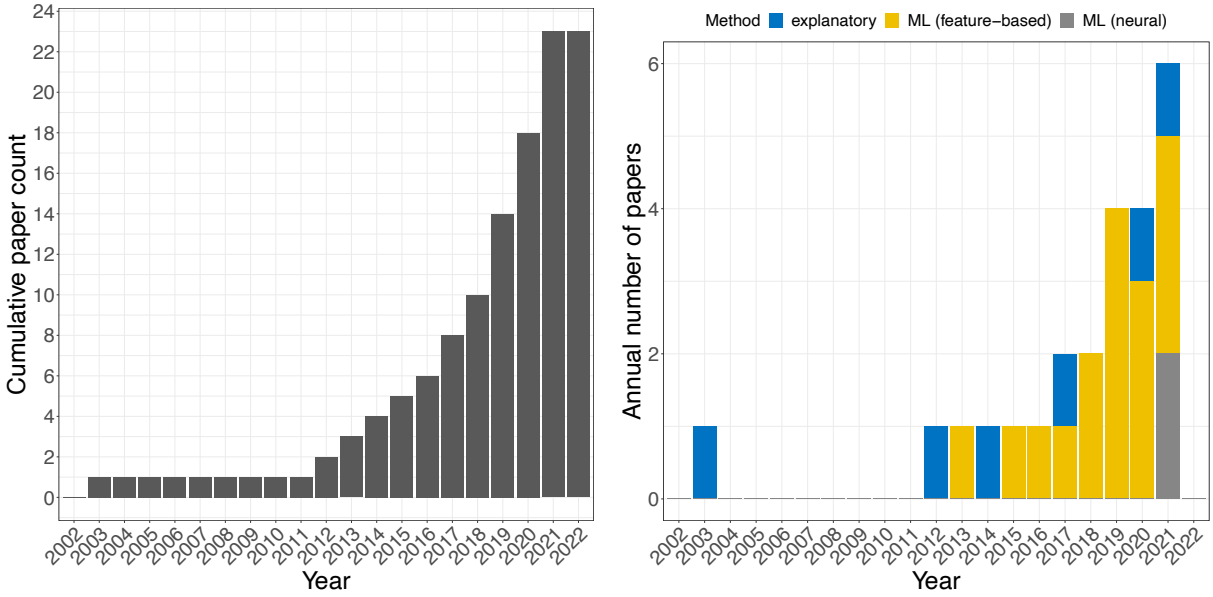
Figure 4.1: *PRISMA flow diagram of the literature identification, screening, and inclusion process for the ALPS survey. The initial records are based on the first 200 hits for each query term. After completing the recommended PRISMA flow, we separated papers that included multiple methodologically unrelated studies relevant for the survey (dashed).*

found a considerable amount of work on ALPS: Additional to 29 surveys on ATS/APA, 81.3% of papers that were assessed for eligibility had to be excluded because they focused on languages other than German not because they were content mismatches. In comparison, only 13.2% of papers were excluded due to our relatively strict inclusion criteria regarding the need for empirical quantitative analyses based on automatically calculated linguistic characteristics. We interpret this combined evidence as an indication that our literature elicitation procedure worked as intended. This supports the hypothesis that there is indeed little work on German ALPS. However, it is possible that some relevant work was not indexed by Google Scholar. We also cannot fully rule out the possibility that a lack of homogeneous terminology in ALPS research partially contributed to the low number of papers. The fragmentation of work on writing quality assessment into a broad range of sub-tasks (which also necessitated the introduction of the term ALPS for this thesis), certainly complicates attempts to gain a comprehensive overview of approaches to the application domain. That being said, we have no reason to believe that these limitations introduced a systematic sampling bias. We can assume that our survey resulted in a representative (if not comprehensive) sample of the relevant literature which is sufficient to address our research questions.

4.2.1 Results

4.2.1.1 Research landscape

Our first research question concerned the different disciplines using ALPS and their statistical methods. Figure 4.2 provides a first overview of the research landscape by showing the longitudinal development of German ALPS research over the past two decades from two perspectives: Figure 4.2a presents a cumulative count to visualize the longitudinal growth of



(a) Cumulative count of papers

(b) Number of new papers per year

Figure 4.2: *Development of the German ALPS research landscape from 2002–2022*

work on ALPS. It includes all papers selected for this survey exactly once ($N = 23$).[7] We see that work on ALPS started as early as 2003 but later work on ALPS did not appear until 2012. From that time on, the number of papers grew annually and increased especially between 2017 and 2021. The seeming lack of growth from 2021 to 2022 can be explained by the literature retrieval date (March 10th, 2022). Overall, this shows that even though work on ALPS was rare in the 2000s, it has attracted increasing research interest in the last decade.

Figure 4.2b displays the number of papers that were published in each year split by the statistical methods they used. We distinguished a) explanatory statistical methods from b) traditional machine learning models using feature engineering and c) end-to-end neural network

[7]The same holds for all subsequent figures in this survey unless explicitly specified otherwise.

approaches. All studies used supervised statistical approaches (i.e. supervised machine learning and supervised explanatory analyses).[8] This is in line with the dominance of supervised machine learning in APA and ATS that I discussed in Section 2.2. Rama and Vajjala (2021) were included twice in this figure because they use both feature-based and neural machine learning ($N = 24$). We identified overall five papers using automatically extracted language indices for explanatory statistical analyses (Strobl, 2014; Vyatkina, 2012; Weiss, 2017a; Daller *et al.*, 2003; Riemenschneider *et al.*, 2021; Ströbel *et al.*, 2020). This includes the two earliest papers found in our survey. Feature-based machine learning approaches to ALPS only started to emerge in 2013 (Hancke, 2013), but have dominated the research landscape since 2015 (Arnold and Weihe, 2016; Frey, 2020b; Rama and Vajjala, 2021; Stiegelmayr and Mieskes, 2018; Szügyi *et al.*, 2019; Vajjala and Rama, 2018; Vanhove *et al.*, 2019; Weiss, 2017b; Bertram *et al.*, 2021; Frey, 2020a; Wahlen *et al.*, 2020; Weiss and Meurers, 2019a,b, 2021; Zesch *et al.*, 2015). End-to-end neural machine learning approaches instead have only started to emerge in 2021. Rama and Vajjala (2021) explored the value of pre-trained embeddings for multilingual (Czech, German, Italian) L2 proficiency assessment, while Ludwig *et al.* (2021) used transformers to score the appropriateness of business e-mails written by vocational and educational training students.

Figure 4.3 shows the research disciplines to which the publication venues of the surveyed papers belong.[9] We distinguished papers based on the statistical methods they used to see potential methodological differences across disciplines. Rama and Vajjala (2021) again contributed two data points. We see that more than half of the papers were published in computational linguistics or computer science venues (CL/CS; $N = 13$). We also found three papers in SLA journals (Vyatkina, 2012; Ströbel *et al.*, 2020; Weiss and Meurers, 2021), two papers in education journals (Bertram *et al.*, 2021; Wahlen *et al.*, 2020), two papers in linguistics journals (Daller *et al.*, 2003; Weiss and Meurers, 2019b), and two papers in interdisciplinary writing research journals (Vanhove *et al.*, 2019; Riemenschneider *et al.*, 2021). One paper each was published in CALL (Strobl, 2014) and psychology (Ludwig *et al.*, 2021) journals. Overall, this confirms that research on or using ALPS for German takes place in a rich interdisciplinary context related to education and is not restricted to computational linguistics and computer science. We also see that exploratory analyses of ALPS hardly take place in CL/CS venues with Weiss (2017a) being the only exception. Machine learning methods instead have

[8] See Section 2.2.3.1 for a more detailed discussion of the different types of approaches.

[9] Theses were attributed to the discipline that is linked to the degree obtained through them.

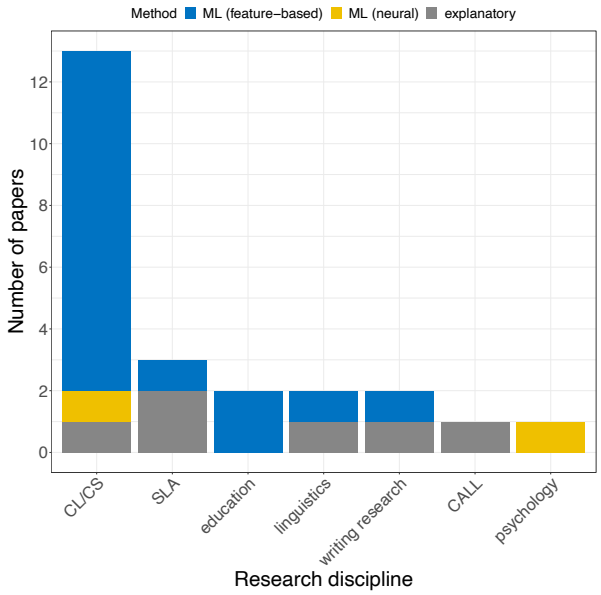
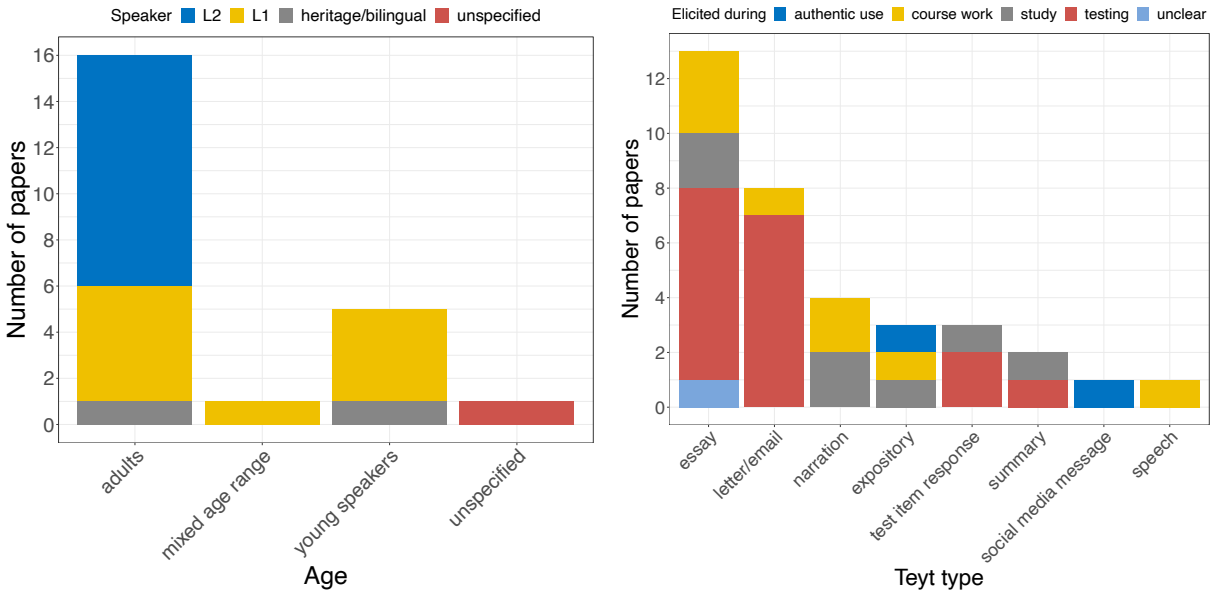


Figure 4.3: *Research disciplines working on or with ALPS split by statistical methods used*

not been limited to publications in computational linguistic or computer science venues, but were also published in journals affiliated with SLA research (Weiss and Meurers, 2021), education research (Bertram *et al.*, 2021; Wahlen *et al.*, 2020), linguistics (Weiss and Meurers, 2019b), interdisciplinary writing research (Vanhove *et al.*, 2019), and psychology (Ludwig *et al.*, 2021). All papers using machine learning techniques trained new, supervised ALPS models. This indicates that models are currently not being re-used after publication.

4.2.1.2 Data sets and labels

Our second research question asked which types of language have been represented in ALPS research and for whom (that is which target groups) ALPS has been conducted. Figure 4.4 addresses this question from two perspectives. Figure 4.4a specifies the target groups for which ALPS approaches were trained. Overall, we see a clear trend towards the analysis of essays written by adults. Most ALPS approaches targeted adults ($N = 16$) either in their L2 ($N = 10$) or L1 ($N = 5$). Work on adults producing L1 German focused nearly exclusively on writing task-specific text quality (Ludwig *et al.*, 2021; Wahlen *et al.*, 2020) or academic writing proficiency (Arnold and Weihe, 2016; Zesch *et al.*, 2015). Only Ströbel *et al.* (2020) analyzed the influence of adults' L1 German writing proficiency on their L2 English writing proficiency. We also found one article concerned with the lexical complexity of spontaneous



(a) Target group specification

(b) Text type and elicitation context

Figure 4.4: *Types of language being targeted by ALPS research*

speech produced by bilingual speakers of German and Turkish (Daller *et al.*, 2003) which was the only paper modeling spoken language performance and one of two papers focusing on bilingual speakers of German.[10] The other paper on bilingual language proficiency was Vanhove *et al.* (2019). They analyzed the proficiency of Portuguese-German bilingual speakers in elementary school. The remaining research on ALPS for children or adolescents focused on L1 writing. Most papers focused on secondary school (Frey, 2020b; Bertram *et al.*, 2021; Riemenschneider *et al.*, 2021; Weiss and Meurers, 2019a). Weiss and Meurers (2019a) focused on both elementary and early secondary school. Frey (2020a) predicted author age and digital familiarity in social media language based on linguistic features. Hers was the only study that covered a mixed age range from children to older adults (aged 10–70+). Stiegelmayr and Mieskes (2018) analyzed web texts with unknown authors. They were marked as fully unspecified in this figure.

Figure 4.4b investigates the types of language that were used in ALPS research by focus-

[10]There might be more work on the assessment of spoken language performance for German that was not retrieved here. Our survey was designed to focus on ALPS for written language as can be seen from our query terms. Thus, we did not systematically elicit studies on spoken language even though we did not exclude them if they fit our inclusion criteria. Future work might extend our survey to also explicitly query for work on spoken language.

ing on the studies' text types and elicitation contexts. Several papers analyzed two to three different text types (Rama and Vajjala, 2021; Szügyi *et al.*, 2019; Vajjala and Rama, 2018; Vanhove *et al.*, 2019; Vyatkina, 2012; Weiss, 2017a,b; Hancke, 2013) and thus contributed multiple data points to the figure ($N = 35$). The most commonly analyzed text type were essays, followed by letters and e-mails. The high number of essays ($N = 13$) and letters ($N = 8$) was partially due to the repeated use of the Merlin corpus (Wisniewski *et al.*, 2013) which includes both text types (Rama and Vajjala, 2021; Szügyi *et al.*, 2019; Vajjala and Rama, 2018; Weiss, 2017b; Hancke, 2013; Weiss and Meurers, 2019b). Other re-occurring text types were expository texts (Arnold and Weihe, 2016; Daller *et al.*, 2003; Ströbel *et al.*, 2020),[11] test item responses (Bertram *et al.*, 2021; Wahlen *et al.*, 2020; Weiss and Meurers, 2021), summaries (Strobl, 2014; Szügyi *et al.*, 2019), and narrative texts. Narrative texts were elicited both for children's writing (Vanhove *et al.*, 2019; Weiss and Meurers, 2019a) and adults (Vyatkina, 2012; Weiss, 2017a). Weiss (2017a) also analyzed drafts for speeches. Frey (2020a) analyzed Facebook messages and posts. This was one of two studies analyzing language that had been produced to fulfill an authentic communicative purpose outside of a learning or testing setting. The second study was Arnold and Weihe (2016), who analyzed on quality differences in Wikipedia articles. Most other language productions were elicited in testing or examination contexts ($N = 17$) or specifically for the participation in a study ($N = 7$). Less research was based on language produced within instructed settings such as language courses (Frey, 2020b; Vyatkina, 2012; Weiss, 2017a; Ströbel *et al.*, 2020).[12] For the web essays elicited by Stiegelmayr and Mieskes (2018), the elicitation context was unclear. Summarizing the findings in Figure 4.4, we see that research on ALPS for German has been targeting many different types of language in terms of target groups and text types. Especially the broad range of text types that has been targeted is promising given the known influence of task effects on language performance (see Section 2.1.3.2). Yet, we could not find any work on young speakers' or writers' L2 proficiency and except for Arnold and Weihe (2016) and Frey (2020a), all work exclusively focused on somewhat artificial language elicitation contexts. More work on language performance elicited in authentic elicitation contexts is needed.

As discussed in Section 4.1, ALPS is an umbrella term coined in this thesis to consider a variety of assessment tasks relating to language performance. This opens the question which types of language performance have been addressed in ALPS research on German. Figure 4.5

[11] This category includes Wikipedia articles, university term papers, and picture descriptions.

[12] Since Vyatkina (2012) and Weiss (2017a) analyzed different text types, they contributed multiple data points to this category.

compares the different types of ALPS tasks that were represented in our survey and the performance scales that they used. Since the survey only retrieved supervised approaches to ALPS, all studies relied on some form of pre-annotated performance scales that could serve as reference labels. Some papers trained multiple classifiers using different scales (Frey, 2020b; Rama and Vajjala, 2021; Frey, 2020a; Hancke, 2013; Riemenschneider *et al.*, 2021; Weiss and Meurers, 2019a), e.g., to compare the performance for different degrees of scoring granularity. This lead to a higher number of data points contributing to this figure ($N = 35$). Most

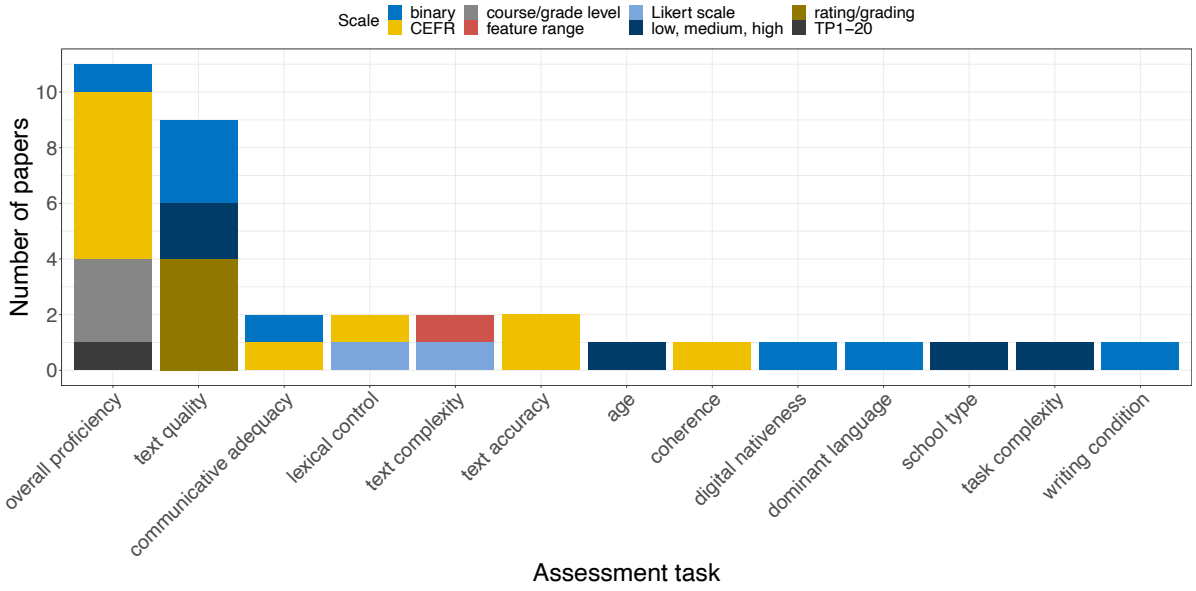


Figure 4.5: *Distribution of ALPS tasks split by performance scales used*

studies estimated overall language proficiency ($N = 11$). The most common scale for this was the CEFR scale from A1 to C1 or C2 (Rama and Vajjala, 2021; Szügyi *et al.*, 2019; Vajjala and Rama, 2018; Weiss, 2017b; Hancke, 2013; Weiss and Meurers, 2019b). Hancke (2013) additionally predicted language proficiency using a binary estimate. It indicated if the CEFR score that learners received was at or above the level of the language test for which they had produced their texts. Large, rated L2 corpora for German are rare, but they are indispensable for supervised machine learning approaches which require reference labels to learn from. All six studies mentioned here used the Merlin corpus (Wisniewski *et al.*, 2013) which provides expert ratings on the CEFR scale (see Section 4.2.1.4 for details). To circumvent the lack of rated L2 corpora for German and analyze language from other data sets, other studies used development in instructed settings as a proxy for proficiency. Vyatkina (2012) tracked L2 de-

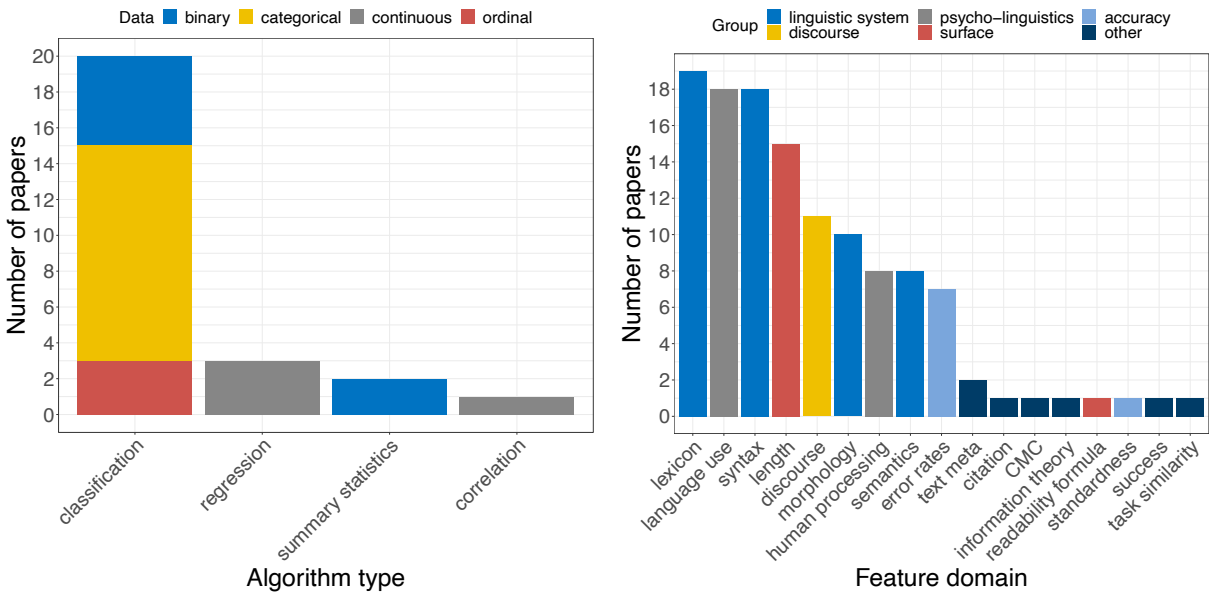
velopment based on elicitation time points during an extended period of L2 instruction in her longitudinal study. Weiss (2017a) and Weiss and Meurers (2021) instead used course levels for the approximation of L2 proficiency in their cross-sectional data sets. Similarly, Weiss and Meurers (2019a) used grade levels for the assessment of L1 academic language proficiency in cross-sectional data.

The second most common ALPS task was the assessment of text quality ($N = 9$) using some form of rating or grading scale (Frey, 2020b; Stiegelmayr and Mieskes, 2018; Riemenschneider *et al.*, 2021; Zesch *et al.*, 2015). Frey (2020b) also used a binary estimate of text quality to predict whether essays were passed or failed and whether they were good or very good essays. Arnold and Weihe (2016), too, used a binary distinction to predict whether Wikipedia articles were featured or not.[13] Unlike grade scales, binary labels were also used for a variety of other assessment tasks including the identification of digital natives (Frey, 2020a), bilingual speakers with German as their dominant language (Daller *et al.*, 2003), and the distinction of collaborative from individual writing (Strobl, 2014). Frey (2020b) and Wahlen *et al.* (2020) used a three-way distinction for the assessment of text quality, which was also used for the prediction of author age (Frey, 2020a), school type (Weiss and Meurers, 2019a), and task complexity (Bertram *et al.*, 2021). Beyond these two main ALPS tasks, our intentionally broad notion of ALPS allowed us to observe a variety of less often represented applications. Both Ludwig *et al.* (2021) and Rama and Vajjala (2021) assessed the communicative adequacy of writing, Ludwig *et al.* (2021) in form of a binary distinction for the classification of professional e-mails and Rama and Vajjala (2021) using the expert ratings on the CEFR scale for the L2 writing in the Merlin corpus. Rama and Vajjala (2021) also used the expert ratings for lexical control, text accuracy, and coherence to locate L2 learners' performance on these subtasks of language proficiency assessment on the CEFR scale. As an alternative to such expert ratings on the CEFR scale (which as mentioned before are rare in German data sets), Vanhove *et al.* (2019) used a Likert scale to measure lexical control. Similarly, Riemenschneider *et al.* (2021) used a Likert scale to measure text complexity. Ströbel *et al.* (2020) instead measured text complexity using complexity feature values elicited in different writing conditions. We saw from this overview that text quality and overall language proficiency assessment until now were the two dominating tasks in research on German ALPS, but that overall a broad range of different tasks has been represented.

[13] On Wikipedia, featured articles are articles that were identified as especially high quality by the Wikipedia editors, see `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`.

4.2.1.3 Methods and evaluation metrics

Our third research question asked which machine learning methods and evaluation metrics have been used in ALPS research. We already approached the subject of machine learning methods in Section 4.2.1.1 by distinguishing publications based on their use of explanatory statistics, feature-based machine learning, or end-to-end neural machine learning. Figure 4.6 elaborates on this further by focusing on the specific methods and feature types that have been used. Figure 4.6a shows how ALPS has been approached statistically across papers for



(a) Prediction methods

(b) Linguistic features

Figure 4.6: *Statistical methods and complexity features used in ALPS for German*

different types of data. Frey (2020b,a) and Hancke (2013) contributed two data points each to this figure because they used binary as well as multi-level classification in both studies ($N = 26$). Most research approached ALPS for German as a classification task ($N = 20$). Of these, most papers treated multi-level labels as categorical ($N = 12$). Only three papers represented multi-level prediction labels as ordinal data (Weiss, 2017a,b; Weiss and Meurers, 2021). Since the underlying performance categories have a natural order, they are intrinsically ordinal and not nominal data. This makes ordinal models the more accurate representation (Gutiérrez *et al.*, 2015).[14] Five papers classified binary data (Arnold and Weihe, 2016; Frey,

[14]Methodologically, ordinal models perform a regression and Weiss (2017a,b) in fact use generalized additive regression models. However, we considered them classification approaches for the purposes of this survey

2020b,a; Hancke, 2013; Ludwig *et al.*, 2021). Overall three studies used regression to analyze continuous labels. Of these, one study used regression for machine learning (Vanhove *et al.*, 2019) and two used regression for explanatory analyses (Riemenschneider *et al.*, 2021; Ströbel *et al.*, 2020). The remaining explanatory analyses used summary statistics (Strobl, 2014; Daller *et al.*, 2003) or correlation analyses (Vyatkina, 2012).

Figure 4.6b shows which types of features were used across papers and to which linguistic domains these belong. We categorized feature types and domains based on the terminology proposed in Section 2.1.2 which can deviate from the terminology used in the respective papers. We included all features that were reported, even if they were not measures of linguistic complexity or not calculated automatically.[15] All papers were included for this figure except Ludwig *et al.* (2021) who used an end-to-end neural approach that did not use any feature engineering.[16] Most papers ($N = 20$) used more than one type of feature and therefore contributed multiple data points to this figure ($N = 123$). In line with the traditionally important role of syntactic and lexical complexity features in research on ATS and language proficiency assessment (see Section 2.1.2), lexicon ($N = 19$) and syntax ($N = 18$) were two of the three most frequent domains. Also features of morphological ($N = 10$) and semantic complexity ($N = 8$) were used repeatedly across studies. As for psycho-linguistic measures of complexity, language use features occurred in most studies ($N = 18$) and were as common as syntax features. Again this is in line with the traditional focus of work on ATS as word frequency measures assess relative lexical complexity (see terminological note at the beginning of Section 3.2). We considered n-gram measures to fall into the same category. Also human processing measures occurred in several studies albeit less often ($N = 8$). It is interesting to note that human processing measures were exclusively measured by studies using a version of the complexity analysis system from this thesis (Frey, 2020b; Weiss, 2017b; Bertram *et al.*, 2021; Riemenschneider *et al.*, 2021; Weiss and Meurers, 2019a,b, 2021).[17] Surface measures of text characteristics were measured mostly in terms of length features ($N = 15$), but Zesch *et al.* (2015) also used readability formulas for ATS.[18] Discourse features were used in nearly

because the final labels that they return are not continuous values but discrete, ordered labels.

[15] Some papers used complexity and accuracy measures or combined automatic and manual features.

[16] Rama and Vajjala (2021) were included here because they not only trained an end-to-end neural classifier but also a feature-based classifier.

[17] See Section 2.1.2.7 for a discussion of human processing measures and Table B.20 in Section B.7 for a complete list and definition.

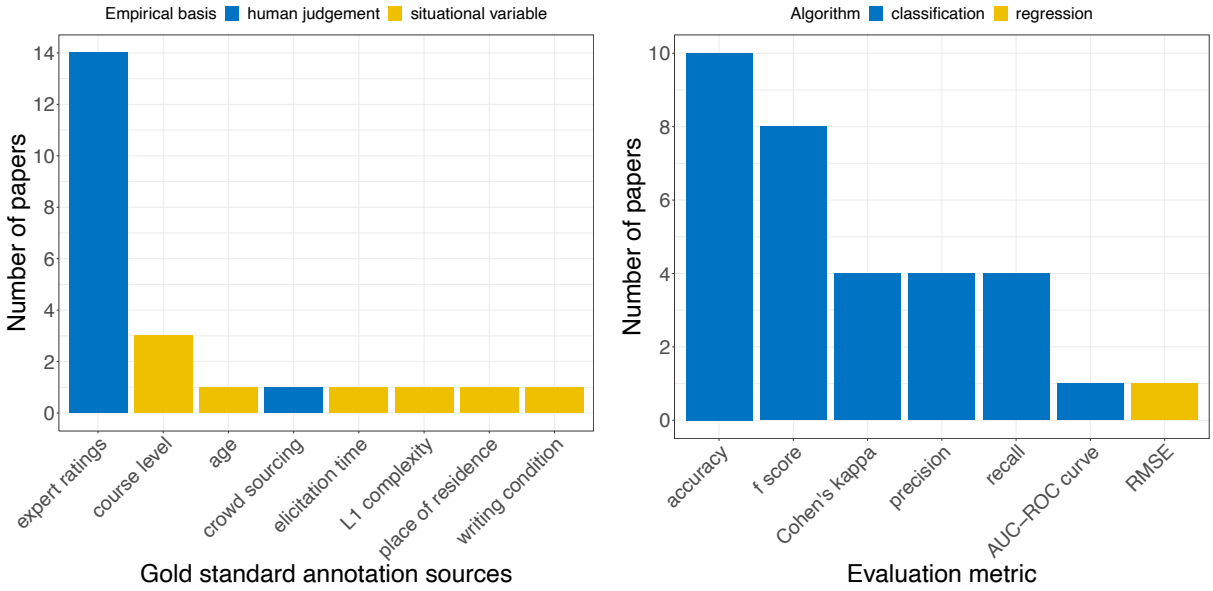
[18] We consider readability formulas as surface level features because they are typically weighted linear combinations of surface level features such as sentence and word length. See Section 2.3.4 for more details.

half the studies included in this survey ($N = 11$). It is surprising to see that they were not used more often, seeing that discourse measures are one of the three central linguistic dimensions in ATS, together with features of lexical and syntactic complexity (cf. Crossley, 2020). Beyond complexity, also measures of accuracy were included in several studies, mostly in form or error rates ($N = 7$) but Frey (2020a) also measured intentional violations against the written standard norm in social media language, e.g., through dialect use or non-standard capitalization or character repetition. Beyond measures of these domains often associated with CAF, we also observed the occasional use of several other types of measures including the presence of citations and quotes (Zesch *et al.*, 2015), the use of computer mediated communication style features (Frey, 2020a), measures from information theory (Ströbel *et al.*, 2020), estimates of input prompt similarity (Zesch *et al.*, 2015), coarse pre-estimates of performance in terms of failure or success (Weiss, 2017b), and meta information on the text type (Vanhove *et al.*, 2019) or task type (Weiss, 2017b) being evaluated. Together, this shows that a broad range of measures have been used for ALPS of German language productions.

Turning to the evaluation of ALPS approaches, Figure 4.7 focuses on estimates of the validity of ALPS reference labels and the performance of predictive ALPS models.[19, 20] Figure 4.7a summarizes the sources of reference labels used for fitting predictive or explanatory ALPS models. This is particularly important seeing that our survey only includes supervised statistical approaches which directly depend on the validity of the reference labels used for training (see Section 2.2.3.1). The majority of studies relied on human judgments ($N = 15$) which were mostly obtained through trained expert raters ($N = 14$). Vanhove *et al.* (2019) instead relied on crowd sourced annotations as reference labels for texts' lexical richness. Together, these two types of human judgments are the source of two third of all reference labels used in studies in this survey. The other third was obtained through various situational variables. Obtaining such situational variables requires much less resources than producing high-quality human judgments for language performances. However, they are much less precise and have a higher risk of lacking construct validity. The most common situational variable is 'temporal progression' estimated in various forms ($N = 4$). Vyatkina (2012) used the time of elicitation in her longitudinal study as reference label. Similarly, Weiss and Meurers (2019a,

[19] See Section 2.2.3.1 (pp. 53–54) for a more detailed discussion of the relevance of model robustness and construct validity.

[20] We are not using the term 'gold standard' here, even though it is the common terminology in machine learning for reference labels that are being used for training or testing models. However, the term implies a certain quality standard and validity that is not met by all of the labels discussed below.



(a) Source of reference labels used

(b) Performance metrics for predictive models

Figure 4.7: *Construct validity of labels and robustness of models in ALPS*

2021) relied on writers' course and grade level as coarse cross-sectional proxy for proficiency assuming that progressing through a series of instructed language courses overall increases learners' language proficiency. Frey (2020a) used age both as proxy for digital nativeness and as variable in itself for her sociolinguistic analysis of age-dependent stylistic differences in language performance in social media. Instead of time as an estimate of development or proficiency, Ströbel *et al.* (2020) used the complexity of L2 learners' L1 writing as an empirical estimate of their general language proficiency. Daller *et al.* (2003) inferred proficiency differences between bilingual German-Turkish speakers based on their place of residence (Turkey or Germany) which they argued to influence which of the two language was the dominant language of bilingual speakers. Finally, Strobl (2014) compared writing performance differences across writing conditions contrasting between collaborative and solitary writing, based on research arguing that collaborative writing improves learners' writing performance. Taken together, this shows that human judgments have been the primary source of reference labels for ALPS for German but that a broad variety of alternative situational variables has been used to substitute the resource intensive labeling process which is a mandatory prerequisite for supervised statistical methods.

Figure 4.7b shows the metrics that were used to evaluate the performance of predictive

ALPS models.[21] Some of the 17 papers used several metrics, increasing the number of data points ($N = 32$). As discussed previously, only Ludwig *et al.* (2021) used regression to train a predictive ALPS model.[22] They evaluated their regression model in terms of RMSE. All other papers training predictive ALPS models used classification. Accuracy ($N = 10$) and f-scores ($N = 8$) were the most commonly used performance metrics and most studies reported them together (Frey, 2020b; Stiegelmayr and Mieskes, 2018; Hancke, 2013; Ludwig *et al.*, 2021; Weiss and Meurers, 2019b). Four of the eight studies reporting f-scores also reported precision and recall as additional metrics (Frey, 2020b; Stiegelmayr and Mieskes, 2018; Weiss, 2017b; Hancke, 2013). Several studies used Cohen's kappa (Cohen, 1960) to evaluate the models' performance (Frey, 2020b; Ludwig *et al.*, 2021; Wahlen *et al.*, 2020; Zesch *et al.*, 2015). This metric of chance-corrected IRR is typically used to estimate the agreement between human annotators. In these papers, the metric was used to compare the automatic predictions and the human reference annotations on the test data. Even though estimates of IRR are not prototypical metrics to evaluate the performance of machine learning systems, this is not an atypical evaluation metric for ALPS. Weighted IRR metrics have been used for different ATS tasks coming from the tradition of assessing (human) rater agreement, see for example the Automated Student Assessment Prize challenge.[23] Only Ludwig *et al.* (2021) also included the AUC-ROC curve, commonly used to evaluate binary models.

Beyond the type of evaluation metric used, the train-test set-up also plays a crucial role for the evaluation of machine learning models. Figure 4.8 summarizes the different types of set-ups used in the 17 machine learning-based studies. Also, six papers used multiple types of test data which increased the total number of data points ($N = 25$). Stiegelmayr and Mieskes (2018) did not specify their train-test set-up within their paper and were labeled as unspecified. Most papers ($N = 15$) used a form of (stratified) n-fold cross-validation (CV) for training and testing. Cross-validation is a common way of training and testing on smaller data sets. Its systematic resampling strategy uses the available data more efficiently. It also reduces the risk of overfitting which is generally elevated with smaller data sets. However, it is computationally more costly than standard train-test splits and can be an issue for larger data sets. The stability of models trained with cross-validation can be tested by comparing the performance variation across folds. In the present survey, none of the fifteen studies reported standard deviations

[21] This analysis treats different varieties of the same metric as identical, e.g., weighted and non-weighted kappa or macro- versus micro-averaged f-scores.

[22] Ludwig *et al.* (2021) used both regression and classification.

[23] `https://www.kaggle.com/competitions/asap-aes/overview/description`

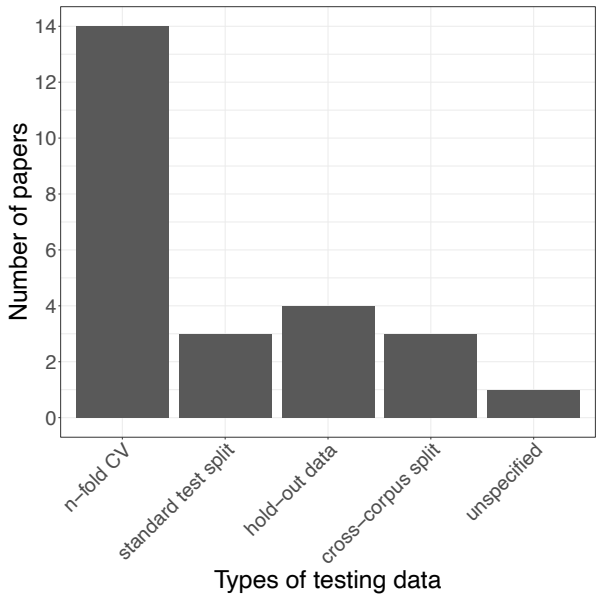


Figure 4.8: *Types of test splits used in machine learning-based ALPS for German*

across folds, though, except Weiss (2017b). A reason for this could be that reporting standard deviations for cross-validation has not (yet) become a common standard in computational linguistic publication outlets. Most of the fifteen studies used ten folds except Rama and Vajjala (2021) who used five folds and Vanhove *et al.* (2019) who used 16 folds. Additionally, Vanhove *et al.* (2019) also tested their model on several standard test data sets which had been split from the full data prior to cross-validation. Ludwig *et al.* (2021) and Weiss and Meurers (2021) were the only ones who did not use cross-validation. They used a standard train-test split instead. Even though the data set used by Ludwig *et al.* (2021) was medium sized ($N = 2,088$), it is possible that computation costs were a concern for their transformer-based approach. However, it is unclear how well their model generalizes to unseen data because they did not have any hold-out or cross-corpus data for additional testing. Weiss and Meurers (2021) trained several models on three larger data sets consisting of answers to reading comprehension questions (between 3,259 and 7,839 data points, see Section 5.2.1.1). They ensured the generalizability of their models through additional evaluations on two types of hold-out data sets (questions and texts) and one cross-subcorpus test set elicited at a different university (see Section 5.2.4). Also some other studies used hold-out and cross-corpus test data. Weiss and Meurers (2019a) tested their models on hold-out writing topics (see Section 5.2.4) and Zesch *et al.* (2015) used hold-out task prompts. Vajjala and Rama (2018) and

Rama and Vajjala (2021) evaluated their models across languages using the subcorpora of the trilingual Merlin corpus (see Section 5.3.4 for a similar approach to cross-language ARA). Despite these encouraging examples, the proportion of studies that has made use of hold-out and cross-copus test splits in our survey is still lower than would be ideal. More research is needed to ensure the generalizability of German ALPS models.

Finally, we investigated if and how the validity of ALPS models has been studied.[24] We found that validation approaches for German ALPS are rare in practice. Only Daller *et al.* (2003) tested the validity of the lexical richness measures that they had used to find performance differences between their subjects. For this, they correlated each score with the C-test results of their subjects as an external measure of language proficiency. They found a statistically significant correlation only for two of their four measures, namely the TTR of for advanced vocabulary and the Guiraud index for advanced vocabulary. No other study in our survey evaluated the validity of their approach.

4.2.1.4 State of the art

Our penultimate research question concerned the current SOTA for predictive ALPS models. We observed that few models are evaluated across corpora and that none of the models discussed here were re-used in later studies. Therefore, we can only evaluate the SOTA in terms of the performance of models on the same corpus. The identification of a SOTA model based on the best cross-corpus generalization is not meaningfully possible. This restricted our assessment to the six studies using the Merlin corpus (Wisniewski *et al.*, 2013) to test a L2 proficiency classifier for German. The performance results are reported in terms of accuracy and weighted F1 score in Table 4.2.[25] There are some differences in the set-up of the six studies, which somewhat limit the comparability of their reported performance metrics. All studies report their performance for predicting the overall proficiency ratings in a five-way classification using the Merlin corpus. However, two studies used marginally different data sets: only Weiss (2017b) and Weiss and Meurers (2019b) included the four C2 texts in the corpus into a joined C1/2 level. The other studies discarded them.[26] Also, most studies used

[24] See Section 2.2.3.1 (pp. 53–54) for a more detailed discussion of the relevance of model validity.

[25] The weighted f-score is particularly suited for unbalanced data sets such as the Merlin corpus (see Section 5.2.1.1), see also the recommendation for the REPROLANG challenge.

[26] It is not possible to learn a statistical representation of the C2 level in the German section of the Merlin corpus because it includes only four C2 texts. Therefore, to use the data for training a model, researchers need to choose a way to address this under-representation, e.g., by excluding the four essays or incorporating C1 and C2 rated essays into a single level.

Table 4.2: *Best performances of ALPS models on Merlin corpus measured by accuracy and weighted f1-score.*

	Levels	Accuracy	F1 score
Szügyi *et al.* (2019)	A1–C1	70.0%	*n.a.*
Hancke (2013)	A1–C1	72.5%	72.4%
Rama and Vajjala (2021)	A1–C1	*n.a.*	69.0%
Vajjala and Rama (2018)	A1–C1	*n.a.*	68.6%
Weiss (2017b)	A1–C1/2	*n.a.*	72.2% / 85.0%
Weiss and Meurers (2019b)	A1–C1/2	70.0%	68.1%

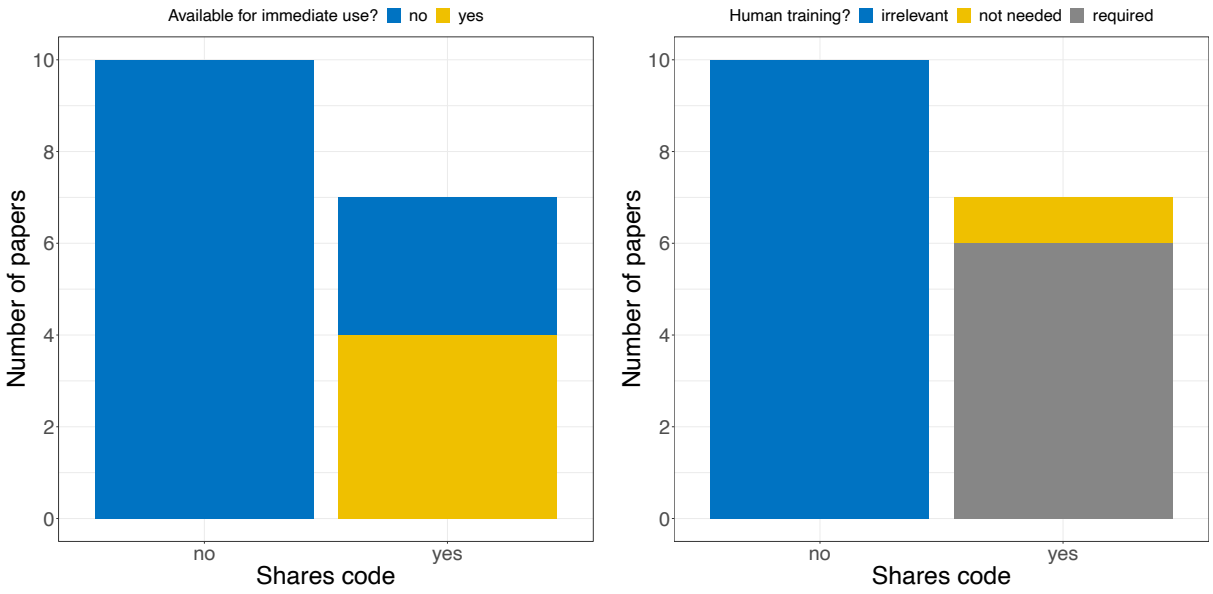
10-CV, except for Rama and Vajjala (2021) who trained and tested using 5-CV, but they used different folds. This might be an issue because except for Weiss (2017b), no one reported estimates for the stability of their models' performance across folds. This limits the comparability of the findings because the scores presented here are the average scores across folds. They do not allow to draw conclusions about the stability of the models. Weiss (2017b) found that the standard deviation of some models' f-scores across folds could be as high as 4%. Assuming that this is representative for the other studies, we cannot estimate whether or not performance differences of at most 4% are indeed significant.

Keeping these limitations in mind, we see that the SOTA f-score for a five-way prediction of overall L2 proficiency on the Merlin corpus lies at 85.0% (Weiss, 2017b). Weiss (2017b) obtained this score for a model that uses not only linguistic complexity features (for which the performance lies at 72.4%) as features but also the information whether or not a learner text was rated at or above the test level (success) or below the test level (failure) at which it was elicited. Unlike the other models, it requires an initial (albeit maximally coarse) performance estimate (namely ±passed). This makes it quite limited for practical use cases. Thus, it is open to debate whether or not we would like to consider this model in our evaluation. The other models instead do not require such an initial estimate. If the model requiring this initial performance estimate is excluded, Hancke (2013) reports the highest accuracy and f-score, closely followed by Weiss (2017b) who also trained a model without the performance estimate as feature. That being said, all models only based on text features are relatively close together in performance falling around an f-score of $70 \pm 2\%$. Assuming that the standard deviations across folds reported by Weiss (2017b) are representative, it is unlikely that these are indeed significant difference. A more conservative estimate would therefore put the SOTA performance of predicting holistic L2 proficiency on the Merlin corpus using a five-way clas-

sification around 70% for both weighted f-score and accuracy.

4.2.1.5 Availability and accessibility of approaches

Finally, we investigated the availability of predictive models for ALPS ($N = 17$). ALPS is of interest for a variety of use cases and users with various degrees of computational and statistical knowledge and skills (see also Section 2.2). Figure 4.9 approaches our final research question from two perspectives. Figure 4.9a identifies papers that shared their trained model



(a) Immediate availability of code

(b) Skill requirements for use

Figure 4.9: *Accessibility of ALPS models*

or (some of) the code or tools needed to directly replicate their study. Most papers ($N = 10$) did not make any resources (models, tools, or code) available, but 27.3% shared the tools or models used in their studies at least partially ($N = 7$).[27] However, hardly any papers shared their resources in a way that made their models, code, or tools accessible to users without the technical skills to replicate machine learning studies or run programming scripts. The only exception to this was Wahlen *et al.* (2020). They used the publicly available ESCRITO toolkit (Zesch and Horbach, 2018) for their prediction of teachers' content knowledge and announced to make their data available for research. Even theough they did not share their trained model,

[27]This also counts studies using freely available tools or systems for their analyses.

this makes their study replicable because users can train the same model on the available data. Since ESCRITO was designed as an end-to-end scoring tool that is also accessible to teaching practitioners, no special training is needed to replicate the study of Wahlen *et al.* (2020). Most other studies for which resources to replicate the study or directly run their models were available made their data, scripts, and results publicly available on OSF (Vanhove *et al.*, 2019) or GitHub (Rama and Vajjala, 2021; Vajjala and Rama, 2018).[28] Some studies make their scripts available, but not their data which makes them not immediately applicable (Ludwig *et al.*, 2021; Weiss and Meurers, 2019a,b). All of these studies have in common that they are not accessible without prior knowledge on how to read or run scripts in R or python. This leaves Wahlen *et al.* (2020) the only paper with a study that is also accessible for users without programming or machine learning skills thanks to the accessibility of the ESCRITO toolkit. This is visualized in Figure 4.9b.

4.2.2 Discussion

Our survey focused on providing a birds-eye snapshot of the last two decades of research on and with ALPS for German. When looking at the history of the research landscape, we saw that research on ALPS for German surfaced as early as 2003, but only really started to be systematically done in 2012. Our survey confirmed that a broad range of research disciplines related to education works with and on ALPS and that the dominant feature-based machine learning approaches to ALPS are not restricted to computational linguistic and computer science research. However, models seem to be rarely re-used after being published.

We also observed that German ALPS mostly focused on the assessment of text quality and overall proficiency for adults writing in their L1 or L2, followed by children and adolescents writing in their L1. Nearly all text data used to fit ALPS models was elicited in (semi-)artificial elicitation contexts. Hardly any to no work focused on children or adolescents writing L2 German and language produced for an authentic communicative purpose (but see Arnold and Weihe, 2016; Frey, 2020a).

In terms of the machine learning methods and features used in ALPS, we saw that categorical and binary classifications dominated. These were being evaluated mostly using accuracy and f-scores. Models were predominantly trained on reference annotations provided through human judgments, but we also observed a variety of situational variables as less resource intensive proxies for performance labels. Future research might benefit from focusing more

[28] All links can be found in the respective papers.

on crowd sourcing ratings, which are less resource intensive than expert ratings but facilitate a more controlled and fine-grained performance estimate than situational variables such as course level or author age, which typically are much more coarse grained and less controlled. Crowd sourcing might also be an interesting way to obtain validations for ALPS models. Currently, validating models' predictions played virtually no role in ALPS research (but see Daller *et al.*, 2003). This is a research gap, that needs to be addressed before ALPS for German can be used responsibly in practice. Similarly, even though we found some papers using hold-out and cross-corpus testing to corroborate the generalizability of their models to new data, most work still relies solely on cross-validation. Future work should make sure to address this shortage. With this desideratum also comes the call for more sizable learner corpora for German with annotations that can be used for supervised machine learning-based ALPS.

Most studies used features of lexical and syntactic complexity, language use, and surface length, in line with the long standing tradition of these features in work on proficiency assessment and ATS. Discourse features were unexpectedly rare, though, considering their importance in ATS for English. One possible explanation for this might be that until recently, there was no tool available for German that facilitated the automatic analysis of discourse features, whereas for English systems such as Coh-Metrix (Graesser *et al.*, 2004) have been available for nearly two decades. In this thesis, this issue is being addressed by extending the CTAP platform to German (see Section 3.3). In fact, seven out of the ten studies measuring discourse features were measured using the CTAP platform or its predecessor pipeline for German (see Section 3.2). We made similar observations for measures of human language processing. This gives reason to believe that future work on ALPS might be able to use a broader range of text features to automatically characterize language performance in German.

Due to the use of different corpora and data sets and the lack of cross-corpus evaluations, we could only consider six papers focusing on automatic proficiency assessment for our assessment of the state-of-the-art in German ALPS. Due to the differences in their study set-up and reporting of results, we were only able to provide a loose characterization of the current SOTA which places the performance around a weighted f-score and accuracy of 70% for models using only text characteristics and 85% for a model including a binary estimate of whether or not learners performed at or above the level that they were tested for or below it. However, we also saw that approaches were not being made available in the majority of studies training ALPS models. This might partially explain the lack of studies re-using ALPS models. The

fact that until now work on German ALPS has focused more on training new models rather than maintaining them, testing their cross-corpus generalization, and making them accessible severely limits the applicability of the SOTA to practice. That being said, a notable number of studies did share their code albeit not in a way that is accessible to users without prior knowledge in machine learning or programming. This shows that ALPS is on a good way to making its resources available, but we need to pursue this further and especially start focusing on enhancing the accessibility outside the computational domain.

Finally, this survey highlighted the immediate contributions of this thesis to German ALPS research. More than a fifth of studies included in the survey are a part of this thesis (Weiss and Meurers, 2019a,b, 2021) or directly based on the work presented here (Bertram *et al.*, 2021; Riemenschneider *et al.*, 2021, see also Section 6.2). Also, 39.1% of the studies included in the survey used one of the two complexity analysis systems presented in Section 3.1 to automatically assess language performance (Frey, 2020b; Weiss, 2017a,b; Bertram *et al.*, 2021; Hancke, 2013; Riemenschneider *et al.*, 2021; Weiss and Meurers, 2019a,b, 2021). This highlights the importance of making the automatic linguistic complexity analysis of German proposed in this thesis publicly available through the CTAP web platform.

4.3 Automatic readability assessment for German: a structured survey of research from 2002 to 2022

The structured ARA survey followed the general study design outlined in Section 4.1. The search terms and patterns used for the literature retrieval script are displayed in Table 4.3. These search terms were identified as the central terms for ARA during the literature research for the narrative literature review reported in Section 2.3. Figure 4.10 shows the individual steps of the literature screening process. The literature identification resulted in overall 2,291 candidate records.[29] Of these, 357 were removed prior to screening because they were duplicates and 34 papers could not be retrieved. The remaining 1,900 papers were screened for their adherence to the inclusion criteria based on their titles, abstracts, or full text. This way, we first considered title and abstract to determine their suitability for the survey. We removed five records because they were written in a language other than English or German. We also removed 1,429 records because they were not concerned with readability assessment and 22

[29]The raw candidate records are available in JSON format in the online supplementary material (`https://osf.io/5vb2x/?view_only=6d1bb8ccfe3f458c946ff4fd6ef5206b`).

Table 4.3: *Google Scholar search terms and patterns used for the structured literature survey of ARA approaches for German (2002-2022, search terms are separated by comma). All search terms are based on the literature commonly used for ARA which was identified in the context of preparing the background chapter on ARA (Section 2.3).*

Language	Search pattern	Search terms
English	"`search term`" AND German	*readability assessment*, *readability formula*, *text readability*, *text assessment*, *text accessibility*, *text difficulty*, *text complexity*, *complexity index*, *Hohenheim comprehensibility index*
German	"`search term`"	*Lesbarkeitsformel* (engl. "readability formula"), *Lesbarkeit* (engl. "readability"), *lesen Komplexitätsindex* (engl. "to read complexity index"), *lesen Verständlichkeit* (engl. "to read comprehensibility"), *lesen Textmerkmale* (engl. "to read text characteristics"), *Hohenheimer Lesbarkeitsindex* (engl. "Hohenheim comprehensibility index")

surveys on readability assessment. We controlled the full texts for the remaining 466 papers. Of these, 262 papers discussed ARA for a language other than German and 72 papers focused on readability assessment for German but violated one of our remaining inclusion criteria. We encoded the remaining 103 studies along the dimensions discussed in Section 4.1.

4.3.1 Results

4.3.1.1 Research landscape

Our first research question focused on the research disciplines and statistical methods used across disciplines for German ARA. To gain an overview of the research landscape and its development in the past 20 years, we first investigated how research on ARA developed in this time period. For this, we focused especially on the comparison of machine learning-based and readability formula-based approaches to ARA. Figure 4.11 tracks the publication of work on ARA over time. Figure 4.11a shows the cumulative growth of papers from 2002 to 2022. It includes all papers considered in this survey exactly once ($N = 103$).[30] We see that starting from the mid 2000s, there has been a systematic increase in ARA publications with increasing jumps in 2017, 2019, and 2021. The seeming lack of growth from 2021 to 2022

[30]The same holds for all subsequent figures in this survey unless explicitly specified otherwise.

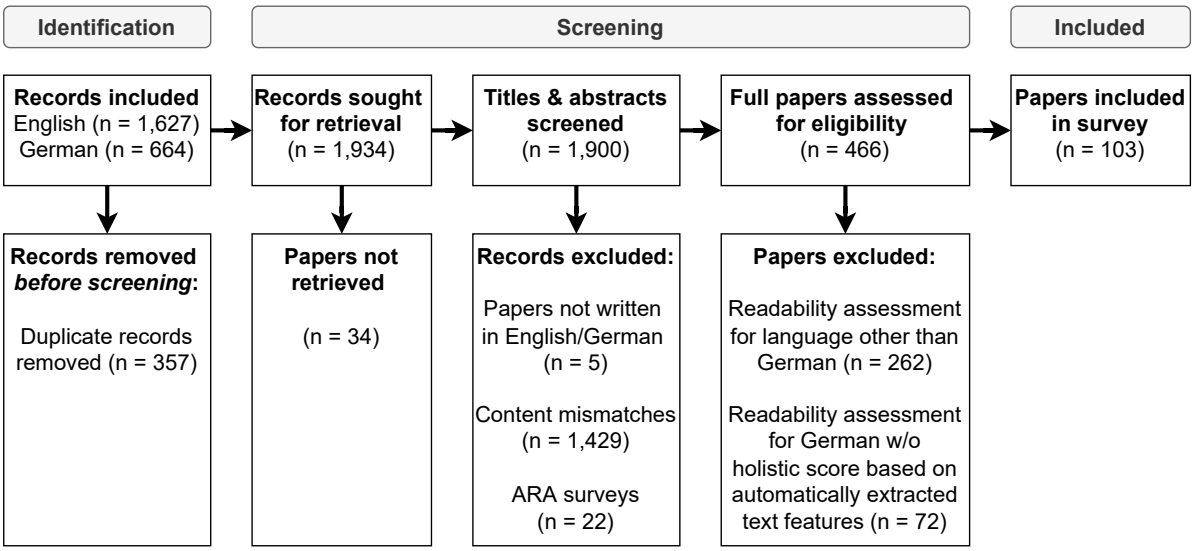


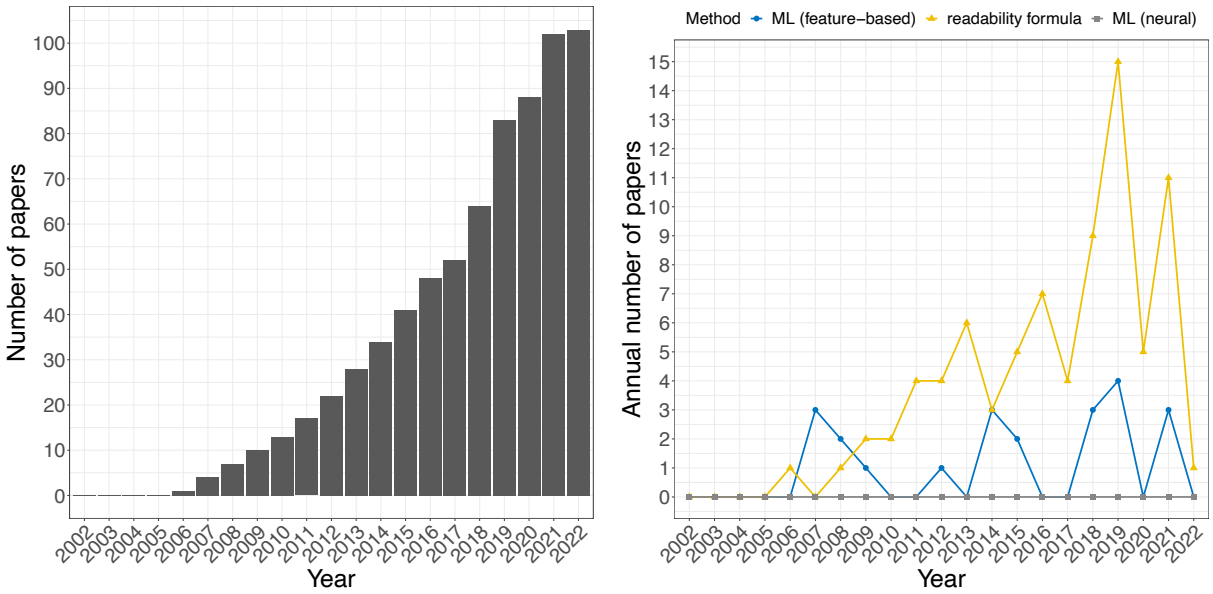
Figure 4.10: *PRISMA flow diagram of literature identification, screening, and inclusion process for ARA survey (initial records based on first 200 hits per query term).*

can be explained by the literature retrieval date (February 16th, 2022).

Figure 4.11b shows the number of papers published each year and differentiates between research based on readability formulas,[31] feature-based machine learning models, and end-to-end neural machine learning models. One paper was excluded because it used an approach that falls in neither of the three categories (Oelke *et al.*, 2010, $N = 102$).[32] Most research on ARA has relied on traditional readability formulas. Feature-based machine learning approaches continue to play a lesser role and neural end-to-end approaches were completely absent in this survey. The earliest automatic, holistic approach to German readability assessment was published in 2006 (Krekeler, 2006) and used the Flesch readability index (Flesch, 1948) for English to control the comprehensibility of German educational materials in a study on the role of background knowledge for L2 readers. From 2008 on, every year new work on German ARA using traditional readability formulas surfaced reaching a peak in 2019. The latest paper using readability formulas for German ARA was a pre-print from the beginning of 2022 which used the Amstad readability index (Amstad, 1978) to showcase the readability differences between simplified and regular texts in a newly introduced corpus for text sim-

[31] Readability formulas are simple estimates of readability based on surface level text characteristics (such as sentence and word length). They have been used since the early 20th century and are still widely despite heavy criticism. For more details, see DuBay (2004, 2006) or Section 2.3.4.

[32] Oelke *et al.* (2010) proposed a tool for visual readability analyses that aggregates individual text characteristics into a holistic score by averaging rather than using learned feature weights.



(a) Cumulative count of ARA papers

(b) Number of new papers per year

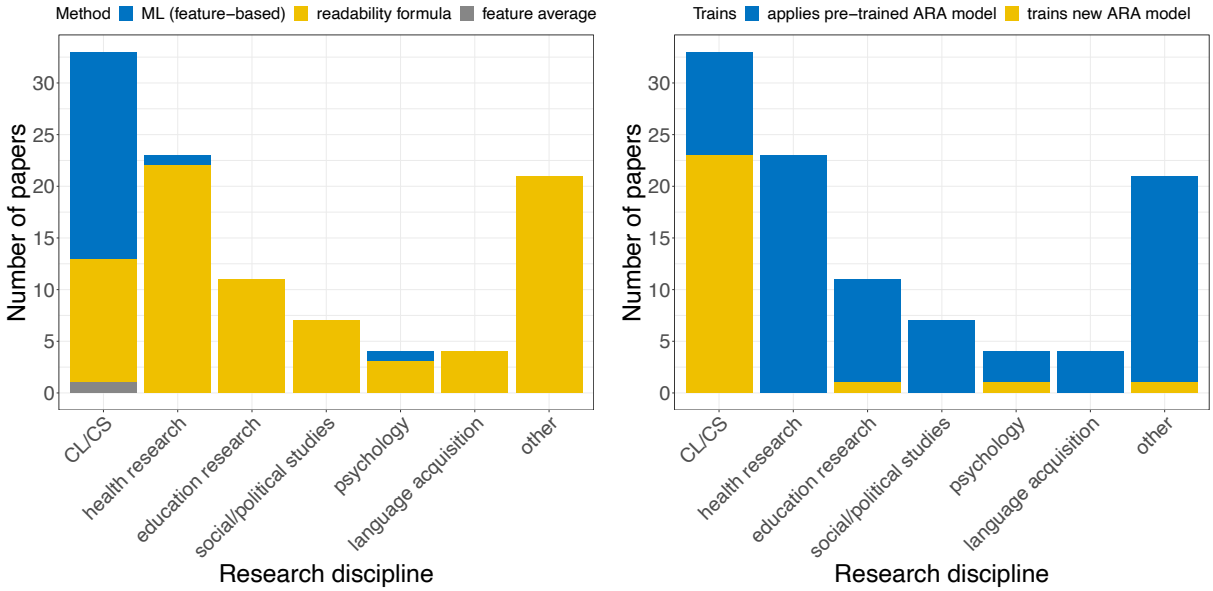
Figure 4.11: *Development of the ARA research landscape from 2002–2022*

plification (Aumiller and Gertz, 2022). Work on feature-based ARA for German started to emerge in 2007. Between 2007 and 2009, the DeLite system (Glöckner *et al.*, 2006) seems to have been the only source of feature-based ARA for German (Brück *et al.*, 2008b; vor der Brück, 2009; Brück and Hartrumpf, 2007a,b; Brück and Leveling, 2007; Brück *et al.*, 2008). This was followed by a research gap until 2012. Hancke *et al.* (2012) used their complexity analysis system for German to train a binary classifier on the first GEO/GEOlino corpus.[33] From that time on, research on feature-based ARA for German increased but has remained less common in our survey than work using readability formulas. Our survey did not retrieve any work using deep learning to predict readability for German, indicating that to date, deep learning has not not played a relevant role for German ARA.[34]

Figure 4.12 explores in which disciplines the different methods were applied (Figure 4.12a) and in which disciplines new ARA approaches have been trained (Figure 4.12b). Even though most papers in this survey came from the areas of computational linguistics and computer

[33] For a more detailed discussion of this corpus, see Section 5.3.1.1.

[34] A possible explanation for this is that the use of end-to-end neural machine learning for ARA in general is a relatively recent development (Vajjala, 2022). The methodological transfer to languages other than English within a computational linguistic research topic itself takes time, as we saw for example in our ALPS survey. Future work should assess how the role of deep learning for German ARA develops within the next years.



(a) Predictive methods used

(b) Training of new ARA models

Figure 4.12: *Comparison of research disciplines working on or with ARA*

science (CL/CS; $N = 33$), we can see that research on and with ARA for German has been published across a broad range of disciplines including but not limited to medicine and health research ($N = 23$), educational research ($N = 11$), social studies and political science ($N = 7$), psychology ($N = 4$), and language acquisition and assessment ($N = 4$).[35] It is striking that machine learning methods have hardly been adopted outside of computational linguistics and computer science, as highlighted in Figure 4.12a. Only two papers published outside of computational linguistics venues made use of machine learning-based classifiers. Berendes *et al.* (2018) trained their own feature-based readability classifier to assess the readability of school textbooks in an interdisciplinary collaboration between computational linguistics and psychology published in a psychology journal. Keinki *et al.* (2018) analyzed the readability of patient information booklets for cancer patients using a binary classification model that had been trained by Zowalla *et al.* (2014) to distinguish medical articles written for experts from medical articles written for laypeople using lexical complexity features. The other studies relied on traditional readability formulas. Also a considerable proportion of papers published in computational linguistics and computer science venues used readability formulas ($N =$

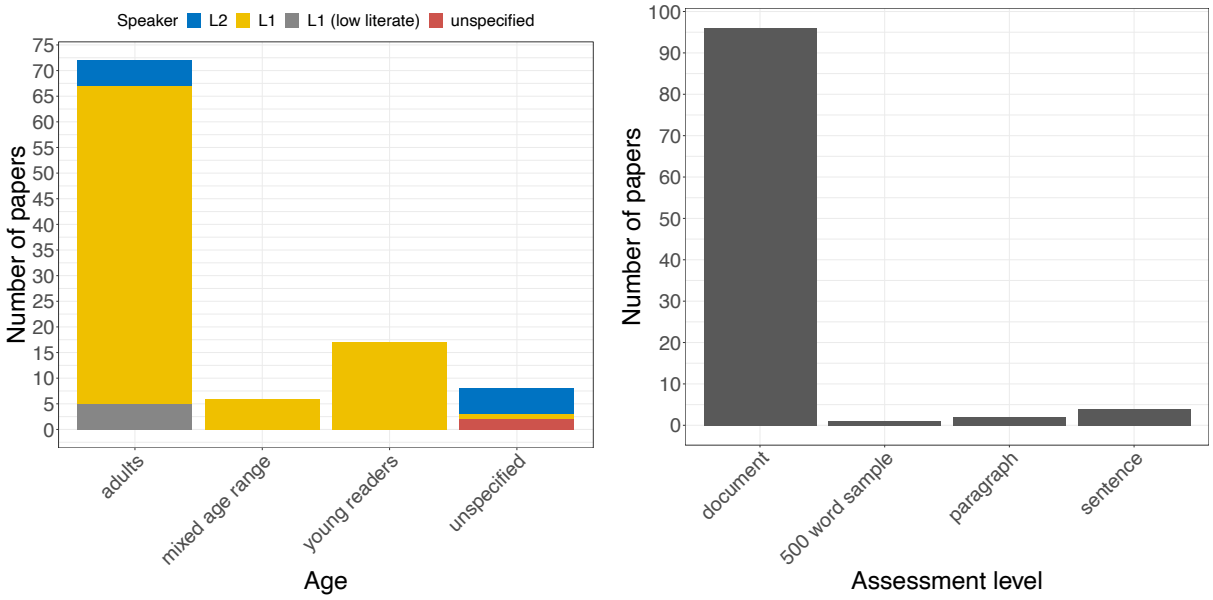
[35] Research domains were identified based on the affiliation of the publication outlet. Theses were linked to a specific research domain based on the department or faculty that they were handed in to. Domains with less than four papers were grouped into the category "other" ($N = 21$).

12). This emphasizes the ongoing practical importance of readability formulas outside of computational research on ARA.

Similarly, Figure 4.12b shows that new ARA approaches were near-exclusively introduced in computational linguistic and computer science papers. There were three notable exceptions to this rule. One was the already mentioned work by Berendes *et al.* (2018) which trained a new feature-based ARA classifier. The other two exceptions proposed new readability formulas, one for young readers (Brügelmann and Brinkmann, 2021) and one for the evaluation of political language (Kercher, 2011). Combining the evidence from both sub-figures in Figure 4.12, we can see that readability formulas are being systematically re-used whereas machine learning-based models are near exclusively used in the studies that introduce them.

4.3.1.2 Data sets and labels

The second focus of this survey lied on determining for which target audiences readability research in German was conducted and which scales were used for this purpose. Figure 4.13a



(a) Target audience specification

(b) Text analysis unit

Figure 4.13: *Types of language being targeted by ARA research*

specifies for which target audiences ARA approaches have been used, differentiating between age groups (adults, young readers, or mixed age range from children to adults) and language

skills (L1, L2, low literate L1). All papers were categorized based on the target group specifications provided within the papers. If the intended target group was not explicitly stated, an approach was labeled as "unspecified" unless the affiliation of target readers could be clearly inferred from the data being analyzed.[36] We see that work on ARA for German predominantly focused on native speakers ($N = 85$) and on adults ($N = 71$). Low literate L1 readers ($N = 5$) are a subgroup in the intersection of these two. There was also some work on L2 readers ($N = 10$) and on L1 readability for young readers ($N = 17$) or mixed age approaches ($N = 6$). However, these domains have been disproportionally under-researched compared to work on adult L1 readers.

Figure 4.13b focuses on the different text levels for which ARA has been conducted. We see that work on ARA for German mostly applied to the document level ($N = 96$). There was also some work at the sentence level ($N = 4$).[37] However, only one of these studies aimed to train an ARA classifier for sentence-wise readability assessment (Naderi *et al.*, 2019a).[38] The remaining papers focused on readability visualization (Oelke *et al.*, 2010), the assessment of reading items (Radner *et al.*, 2016), and reading performance evaluation (Nagler *et al.*, 2014). Work on other text units has been negligible (Kefer, 2013; Klas, 2011; Locher *et al.*, 2019).

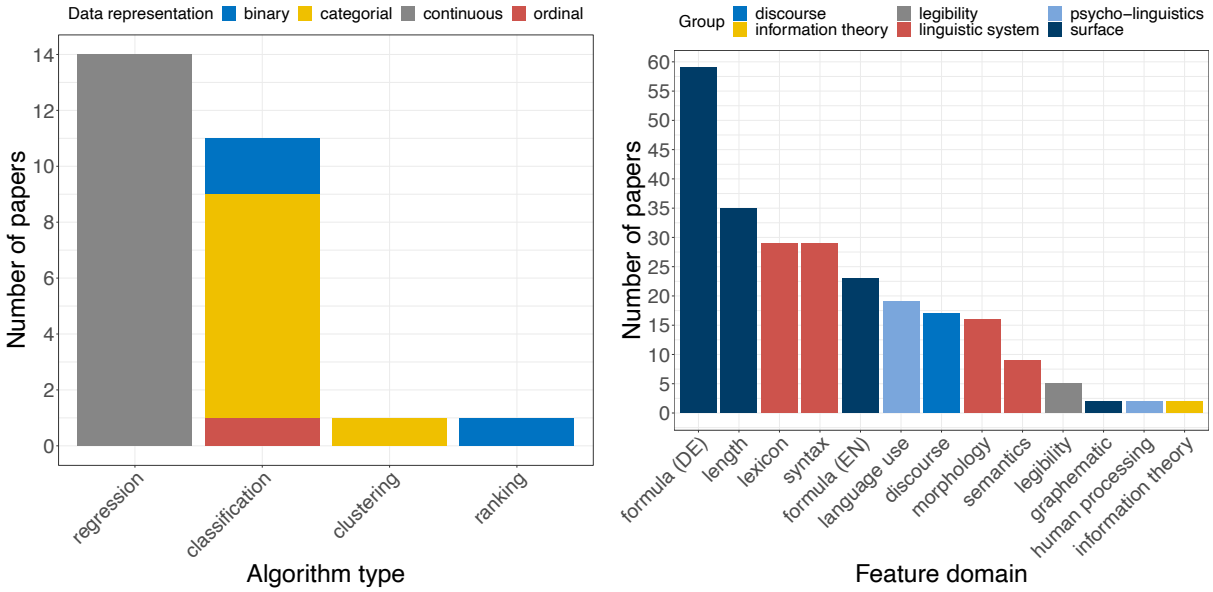
4.3.1.3 Methods and evaluation metrics

Our third research question focused on the different statistical approaches prevalent in work on ARA for German and how they have been evaluated and validated. Figure 4.14 shows the different types of predictive statistical models and complexity features that were used for German ARA. Figure 4.14a investigates which newly trained or fitted predictive statistical modeling approaches (regression, classification, ranking, clustering) were used to represent reference labels of different data types (binary, categorical, continuous, ordinal). This included overall 27 studies. It can be seen that most papers formulated ARA as a regression problem ($N = 14$). The second most common algorithm type were classification algorithms ($N = 11$). Most of these used a categorical classification approach that is agnostic to the inherent order in readability labels ($N = 8$). Even though ordinal classification is the more accurate representation for this type of data (Gutiérrez *et al.*, 2015), only Weiss *et al.* (2021) used ordinal classifi-

[36] For example: municipal, political, or medical texts—which we systematically assumed to clearly target adult native speakers—or schoolbooks—which we systematically assumed to target young (near-)L1 readers.

[37] Work below the sentence level was excluded from the survey based on our inclusion criteria.

[38] The sentence-wise readability assessment study that is part of this thesis (Weiss and Meurers, 2022) was published after the literature retrieval for ARA papers.



(a) Prediction methods

(b) Linguistic features

Figure 4.14: *Statistical methods and complexity features used in ARA for German*

cation for the distinction of multiple reading categories. Two papers used classification for binary categories (Hewett and Stede, 2021; Weiss and Meurers, 2018). Ranking (Vlachos and Lappas, 2011) and clustering (Battisti *et al.*, 2019) were less commonly used.

Figure 4.14b zooms in on the linguistic domains that were represented across ARA approaches. It includes all papers except Imperial and Ong (2021) who did not make explicit which features they used ($N = 102$). Most papers made use of features from multiple domains leading to them being represented multiple times in this figure ($N = 247$). We see that the group of surface measures were by far most common. This includes German readability formulas ($N = 59$) and surface length measures ($N = 35$), but also English readability formulas ($N = 23$).[39] Purely graphematic features also fell into the category of surface measures but were less common (Battisti, 2019; Battisti *et al.*, 2019). Complexity measures of the linguistic system are the second most distributed feature group. It is dominated by lexical features ($N = 29$) and syntactic features ($N = 29$). Less work employed morphological complexity features ($N = 16$) or features of semantic complexity ($N = 9$). Together, these two groups accounted for 81.8% of feature domains represented in the surveyed ARA research. Outside

[39]Even though most features from English readability formulas can be applied to German texts, it remains unclear how meaningful they are for German. Evidently this does not stop their use in practice.

of these two groups, only discourse features ($N = 17$) and language use features ($N = 19$) were relatively frequent. Studies using legibility features were quite rare ($N = 5$)[40] and little work included measures of human processing (Weiss and Meurers, 2018; Weiss *et al.*, 2021) or measures derived from information theory (Islam, 2014; Budd *et al.*, 2019). This overall demonstrates that despite the dominance of readability formulas (and the concerning application of English formulas for German text data), a broad range of linguistic features has been used in ARA approaches for German.

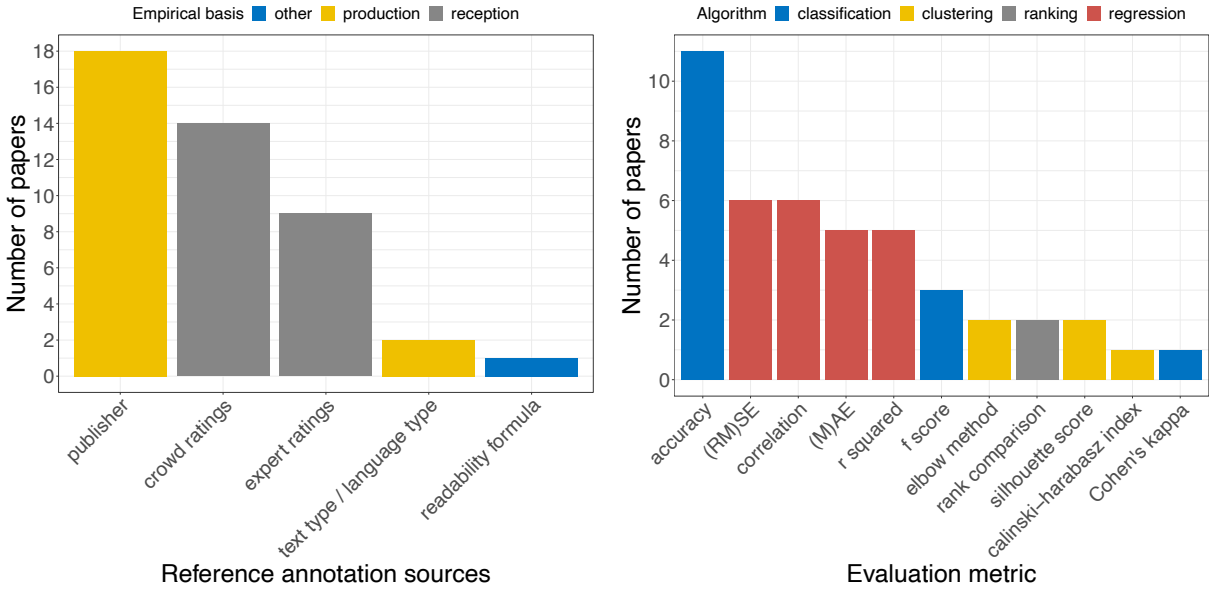
Moving from the methods to evaluation of ARA approaches, Figure 4.15a focused on the empirical validity of reference labels and metrics for model robustness used.[41, 42] It is based on all papers evaluating an ARA approach against a set of reference labels ($N = 43$). Fricke (2021) contributed two data points because he used both, teacher (i.e. expert) and student (i.e. non-experts) judgments. This lead to a total count of 44 data points. The figure distinguishes between different empirical basis for the human annotation of reference labels. Most papers compared their ARA approaches against reference labels that were provided by publishers ($N = 18$). This is a form of production-based reference label in the sense that the texts' producers (e.g., writers or editors) worked towards matching their texts to a predefined proficiency level. However, also reference labels obtained through annotation experiments were common, both in form of crowd ratings, i.e. ratings from a large group of untrained annotators ($N = 14$),[43] or in form of expert ratings ($N = 9$). Both are reception-based labels in the sense that they were assigned after texts had been produced based on the judgments of texts' recipients. As discussed in Section 2.3.3.2, labels by experienced publishers of leveled reading materials can be a reliable and resource efficient source of labeled reading materials. Yet, these labels are typically not independently tested for their ecological validity. Experimentally obtained labels can overcome this limitation but are often not feasible for the annotation of large quantities of data needed for supervised machine learning approaches. Besides these

[40] This survey explicitly excluded papers focusing only on legibility. Thus, the low number of studies including legibility measures is not representative of a lack of studies on these types of features but of a lack of studies combining legibility and readability features.

[41] See Section 2.2.3.1 (pp. 53–54) for a more detailed discussion of the relevance of model robustness and construct validity.

[42] We are not using the term 'gold standard' here, even though it is the common terminology in machine learning for reference labels that are being used for training or testing models. However, the term implies a certain quality standard and validity that is not met by all of the labels discussed below.

[43] Our notion of crowd ratings focuses on their property of compensating annotator training with the majority intuition of a larger group of untrained annotators. We do not require crowd ratings to have been obtained through a crowd-sourcing platform.



(a) Source of reference labels

(b) Model evaluation metrics

Figure 4.15: *Construct validity of labels and robustness of models in ARA*

three most prevalent sources for reference labels, some papers also compared their ARA approaches against text types (e.g., simplified versus non-simplified texts) or genres (e.g., children's books vs. election programs vs. scientific papers) that come with an implied gradient in comprehensibility (Battisti *et al.*, 2019; Oelke *et al.*, 2010). Nagler *et al.* (2014) used the predictions of another readability formula as reference label.

Figure 4.15b displays which evaluation metrics papers used to assess the performance of the ARA models that they used when predicting reference labels. This comparison is based on the 30 papers that evaluated model robustness through estimates of prediction performance or goodness of fit. The figure includes overall 44 data points because 12 papers reported more than one evaluation metric. The figure clearly shows that classification approaches mostly relied on accuracy ($N = 11$). Only three papers reported f-scores additional to accuracy (Islam, 2014; Galasso, 2014; Imperial and Ong, 2021). No paper reported precision and recall for their models. Weiss *et al.* (2018) used Cohen's kappa. In want of data annotated with ecologically validated reference labels for low literate readers, they chose to compare the predictions of their readability model with human annotators.[44] For the evaluation of regression-based ARA models a more varied range of metrics has been utilized. Error metrics were the most

[44]This reasoning is similar to the motivation of using IRR metrics in the evaluation of ATS models.

common choice ($N = 11$), calculated in form of (mean) absolute error (Brück and Hartrumpf, 2007b; Brück and Leveling, 2007), (root mean) squared error (Budd *et al.*, 2019; Naderi *et al.*, 2019a), or both (Brück *et al.*, 2008b; vor der Brück, 2009; Brück and Leveling, 2007; Brück *et al.*, 2008). Also correlation metrics were used often, but mostly in combination with other metrics such as R^2 (Grzybek, 2010; Harbach *et al.*, 2013) or error metrics (Brück *et al.*, 2008b; Brück and Hartrumpf, 2007a,b). Only Brügelmann and Brinkmann (2021) used correlation as sole evaluation metric to compare the predictions of readability formulas that used different scales. Several papers utilized R^2 to evaluate their models in terms of variance explained (Gilg *et al.*, 2019; Grzybek, 2010; Harbach *et al.*, 2013; Merges, 2014; Theobald *et al.*, 2021). Battisti (2019) and Battisti *et al.* (2019) evaluated the quality of their clustering approaches using a combination of elbow method and silhouette score. Battisti *et al.* (2019) additionally calculated the Calinski-Harabasz index (Caliński and Harabasz, 1974). Two papers evaluated their methods in terms of the relative ranking of documents (Kercher, 2010; Vlachos and Lappas, 2011). These findings demonstrate that while there is a broad range of evaluation metrics being used for regression models, classification-based ARA for German could benefit from reporting a more comprehensive selection of evaluation metrics. It would be desirable to report f-scores more systematically and to also include estimates of precision and recall because they provide a more informative assessment of model performance.

To judge the robustness of a new model, it is central to not only consider its performance metrics on their own but also how it was trained and tested. Figure 4.16a shows how common different train-test splits were in the past two decades of research on ARA for German. It includes only papers that proposed a new ARA approach ($N = 31$). Four papers used different types of train-test splits and contributed several data points ($N = 35$). Most studies used a form of cross-validation, typically with ten folds ($N = 10$), but some papers also used seven (Naderi *et al.*, 2019a), five (Budd *et al.*, 2019; Imperial and Ong, 2021), or three folds (Brück and Leveling, 2007). Vlachos and Lappas (2011) and Zowalla *et al.* (2014) did not report the number of folds they used. Two papers used a train-test split (Islam, 2014; Brück and Hartrumpf, 2007a) but they did not control the generalizability of their models through other means. In fact, only four papers investigated the generalizability of their ARA models on some form of hold-out or cross-corpus data. Berendes *et al.* (2018) evaluated the generalizability of their approach using hold-out publisher data. Weiss and Meurers (2018) confirmed the robustness of their readability models across different corpora of educational media language (see Section 5.3.2). Our survey also includes two multi-lingual approaches to ARA that included

(a) Test splits used

(b) Source of reference labels

Figure 4.16: *Comparison of test methods used for ARA models*

German and employed cross-language testing (Budd *et al.*, 2019; Weiss *et al.*, 2021). Not all papers that proposed a new ARA approach required training data. Several papers focused on fitting explanatory regression or factor models (Brügelmann and Brinkmann, 2021; Gilg *et al.*, 2019; Grzybek, 2010; Harbach *et al.*, 2013; Merges, 2014; Theobald *et al.*, 2021; Tolochko and Boomgaarden, 2019; Brück and Hartrumpf, 2007b) without evaluating their predictive power. Also, three papers did not learn new prediction weights based on training data making a train-test split superfluous (Hewett and Stede, 2021; Oelke *et al.*, 2010; Weiss *et al.*, 2018). Similarly, the unsupervised approaches by Battisti (2019); Battisti *et al.* (2019) did not require a train-test split. Overall, this comparison showed that despite some promising examples of cross-corpus and hold-out data testing, most German ARA models are being insufficiently tested with regard to their generalizability to new data sets.

After investigating concerns of construct validity, model robustness, and model generalizability, Figure 4.16b focuses on the assessment of model validity across papers. It considers all papers but includes Fricke (2021) twice because he uses two validation methods ($N = 104$). The figure clearly shows that the vast majority of papers using ARA did not ensure the validity of the predictions ($N = 84$). This is in line with the observations made by Vajjala (2022) for ARA for English. Overall 20 studies included some form of validation based on ecologically

validated observations using mostly human (non-)expert judgments ($N = 12$), but also reading comprehension tests (Beime and Menges, 2012; Golke and Wittwer, 2017; Harbach *et al.*, 2013; Kefer, 2013; Soemer *et al.*, 2019; Tolochko and Boomgaarden, 2019), reading time estimates (Nagler *et al.*, 2014), or EEG experiments (Andreessen *et al.*, 2021). However, half of these studies did not aim to validate the ARA models that they used. They relied on ARA to confirm their assumptions about the reading materials that they used as stimuli in reading studies (Andreessen *et al.*, 2021; Betschart *et al.*, 2019; Dirga and Wijayati, 2018; Friedrich and Heise, 2019; Golke and Wittwer, 2017; Lampert *et al.*, 2016; Nagler *et al.*, 2014; Soemer *et al.*, 2019; Vössing and Stamov-Rossnagel, 2016; Vössing *et al.*, 2016). Of the ten studies that measured the validity of ARA, all but one used established readability formulas for ARA such as LIX (Björnsson, 1983), Amstad's (1978) readability index, or the *Wiener Sachtextformeln* (engl. "Vienna text formulas") by Bamberger and Vanecek (1984). Only Merges (2014) validated the readability formula that he fitted using non-expert judgments. None of the predictive machine-learning based approaches validated the predictions of their models. These findings highlight the need for validating ARA models for German.

4.3.1.4 State of the art

To answer our penultimate research question, we investigated the state-of-the-art performance for ARA models. This typically entails two assessment dimensions: the identification of the current best performance on one or more benchmark data sets and the identification of a model that demonstrated SOTA performances across data sets. However, the number of available readability corpora has been limited. Most machine learning-based studies in this survey used their own data sets which made it difficult to compare the performance of ARA approaches across papers. We also observed that only few studies tested their models across data sets. Thus, this comparison focuses on two corpora which were used to test ARA models across multiple studies: the ReadingDemands corpus (Vajjala, 2015) and the GEO/GEOlino corpus (Hancke *et al.*, 2012; Weiss and Meurers, 2018).[45] The ReadingDemands corpus consists of reading texts from German geography textbooks published for grades five to ten in two types of German secondary schools: *Hauptschule* (engl. "vocational track") and *Gymnasium* (engl. "academic track"). Table 4.4 summarizes the performances of the three ARA models trained on this data. Weiss (2015), Vajjala (2015), and Berendes *et al.* (2018) used this data to train

[45] Weiss and Meurers (2018) proposed an extended and updated version of the corpus originally compiled by Hancke *et al.* (2012). For more details, see Section 5.3.1.1.

Table 4.4: *Accuracy of ARA models on ReadingDemands corpus*

Paper	Grade	School (all)	School (5/6)	School (7/8)	School (9/10)
Berendes *et al.* (2018)	53.7%	*n.a.*	76.8%	78.0%	77.9%
Vajjala (2015)	53.3%	74.5%	*n.a.*	*n.a.*	*n.a.*
Weiss (2015)	53.9%	76.9%	*n.a.*	*n.a.*	*n.a.*

Table 4.5: *Best performances of ARA models on GEO/GEOlino corpus*

	Corpus by Hancke *et al.* (2012)	Corpus by Weiss and Meurers (2018)
Galasso (2014)	89.8%	*n.a.*
Hancke *et al.* (2012)	89.7%	*n.a.*
Imperial and Ong (2021)	*n.a.*	89.0%
Islam (2014)	88.1%	*n.a.*
Weiss and Meurers (2018)	91.1%	89.4%

ARA models that predicted school track (binary) and grade level (5/6, 7/8, 9/10) using 10-CV. All three papers reported similar performances for grade level identification using comparable set-ups. This can be partially explained by the similarity of their set-ups. All used a Sequential Minimal Optimization (SMO) classifier informed by linguistic complexity features (syntax, lexicon, morphology, discourse, language use) that were extracted with the same feature extraction system (an earlier version of the systems used in this thesis). For school type identification, Berendes *et al.* (2018) and Weiss (2015) outperformed Vajjala (2015), leaving the SOTA for this data around 77.0%. Weiss (2015) introduced several new features of lexical and syntactic complexity, which were also used in Berendes *et al.* (2018) but not in Vajjala (2015). This might explain the higher performance of these papers for school types identification. However, Berendes *et al.* (2018) only reported the grade-wise accuracy of predicting school track. This only allows an approximate characterization of the SOTA performance for this data.

Table 4.5 summarizes the best performances on the GEO/GEOlino corpus. It consists of articles from the two popular science magazines GEO (for adults) and GEOlino (for children) which were produced by the same publisher. Overall five papers reported the performance of their ARA classifier on this data.[46] However, they partially reported different corpus sizes

[46] Andreessen *et al.* (2021) were excluded even though they used the same corpus, because they did not evaluate an ARA model on it.

and used either 10-CV (Galasso, 2014; Hancke *et al.*, 2012; Weiss and Meurers, 2018), 5-fold cross-validation (Imperial and Ong, 2021), or a train-test split (Islam, 2014) for their models, which limits the conclusiveness of the performance comparison. With this disclaimer in mind, Weiss and Meurers (2018) reported the highest accuracy on both the original version of the corpus (91.1%) and the new, extended version (89.4%). The model has also demonstrated its cross-corpus generalizability on a comparable corpus of German media language, the Tagesschau/Logo corpus that was compiled by Weiss and Meurers (2018).

4.3.1.5 Availability and accessibility of approaches

Finally, this survey was concerned with the question of model and code availability and accessibility. Benjamin (2012) pointed out the importance of ensuring that ARA methods are accessible to a broad audience so that they can be used in practice and interdisciplinary research. In the last evaluation step of this survey, we investigated which of the 26 papers introducing new ARA methods made their ARA models publicly accessible, which of these were directly usable, and whether or not a technical background was needed to operate them. We found that overall six papers shared their ARA approaches by explicitly stating their formulas including weights or by sharing their data, models, and scripts (Brügelmann and Brinkmann, 2021; Kercher, 2011; Merges, 2014; Weiss *et al.*, 2018, 2021). Three papers made their ARA approaches accessible for immediate use (i.e. do not require users to re-implement the approach to be able to use the analysis): The readability formula by Weiss *et al.* (2018) was integrated into the interface of the search engine `www.kansas-suche.de` into which users may also upload their own texts. Budd *et al.* (2019) integrated their machine learning model into the Auto-ILR system. Both of these are immediately accessible to an audience without the technical skills to re-implement the approach. Weiss *et al.* (2021) made their data, scripts, and trained models accessible online. Their feature-based machine learning approach was informed by features calculated with the freely available CTAP web platform. Thus, their models can be directly used to analyze new data provided that users have the technical skills to run the scripts and load the models.

4.3.2 Discussion

This survey focused on providing a snapshot of the research that has been conducted on and using ARA for German. Over the past two decades, we saw a steady increase of work us-

ing ARA for German or developing new ARA approaches. Remarkably, we saw that ARA approaches seem to have become a tool that is used across a broad range of research disciplines ranging from computational linguistics and education to health research and political science. However, we also observed a clear division of methods: the development of new ARA approaches using state-of-the-art machine learning methods remains the task of research on ARA in computational linguistics and computer science. In other disciplines, new models are rarely proposed and if so they are traditional readability formulas. Outside of computational linguistic work dedicated to developing new ARA approaches, readability formulas remain the de facto standard of ARA for German, even though a) statistical machine learning methods have repeatedly been shown to outperform traditional formulas and b) newly introduced approaches predominantly utilize feature-based machine learning models. We also saw that deep learning approaches have not (yet) reached ARA for German. This has serious implications for ARA research. It means that the current SOTA remains disconnected from research practice let alone everyday practice. As discussed in Section 2.3, the ubiquity of readability formulas in ARA-based research (which we could also observe in this survey for German) can be partially explained by the ease of use and accessibility of formulas. Thus, statistical ARA models must become more accessible. Our survey showed that most papers proposing new ARA methods do not make their models accessible in a way that allows other ARA researchers or even potential users with limited technical skills to utilize them. As a result, new ARA are rarely re-used. The impact of SOTA ARA approaches for German could be greatly strengthened, if the trained models would be made available online and possibly incorporated in user interfaces that facilitate their immediate use for new unlabeled data.

Addressing our second research question, we saw that most ARA research for German focuses on adults and native speakers. Readability is mostly being assessed at the level of full documents. Yet, the research landscape has proven to be quite diversified in terms of the type of language being targeted. It includes work for children, language learners, and shorter text segments. More research in these directions is needed to redress the imbalance, but this seems to be well on its way.

We put a strong emphasis on statistical methods and model evaluation including concerns of the robustness, generalizability, and validity of ARA approaches. We could confirm that regression and classification are the primary ways of framing ARA in statistical approaches, although ranking and clustering have been explored. The robustness of these models is mostly evaluated based on accuracy for classification approaches and different error rates for regres-

sion approaches. Work on classification might benefit from using evaluation metrics more often that penalize classification biases more, e.g., f-scores, precision, and recall. The reference labels used to train these ARA models are mostly assigned by publishers, but also crowd and expert judgments are relatively common. Only few studies tested the validity of model predictions and these were focused on readability formulas. No machine learning-based approach has been tested for its ecological validity. This confirms the need for more work on the validity of ARA models pointed out by Vajjala (2022) for English. Thus, the second foremost research desideratum that emerges from this survey is to increase empirical validation of reference labels and model predictions.

Inspecting the types of features used in feature-based machine learning approaches, we saw that even though a broad range of feature types has been covered, length-based surface features are the most dominant group following the historical precedent of traditional formulas. They are often combined with syntactic and lexical features in line with the dominance of these domains in related complexity research. Surprisingly, morphological complexity features are as common as language use and discourse based features attesting that at least in languages other than English, this domain might not be not as under-researched as often assumed. Graphematic features, measures derived from information theory, and human processing measures, instead, are rarely used. It might be worthwhile to investigate these domains further in the future.

We found two corpora that had been re-used across studies in a way that allowed us to identify the SOTA performance for these data sets. Both represent readability for native speakers. On the GEO/GEOlino corpus distinguishing media language targeting adults and children, the best performing model was presented by Weiss and Meurers (2018) achieving an accuracy of 91.1% on the corpus by Hancke *et al.* (2012) and generalizing well in a cross-corpus evaluation on a comparable corpus (see Section 5.3.2). However, this model performs only a relatively coarse-grained binary distinction. We also compared three models trained on the ReadingDemands corpus (Vajjala, 2015). They all achieved comparable accuracies around 53-54% for the distinction of grade levels (fifth to tenth grade) and around 75% for the distinction of secondary school types. However, no cross-corpus validation studies were reported for these models. Also, none of these models was made publicly available which is in line with our general observation that relatively little machine learning-based work on ARA has been made accessible to other researchers or non-technical audiences. This only highlights the before mentioned need for making sure that machine learning-based ARA models are being

made available to strengthen their impact on ARA practice.

Finally, this survey highlighted several contributions made in this thesis to advance ARA for German. Not only did we see that the model proposed in Weiss and Meurers (2018) seems to remain the SOTA in terms of its performance on the GEO/GEOlino corpus and its cross-corpus generalizability. The survey also highlighted the need for more cross-corpus validation for ARA models. Two out of the four papers that used a form of cross-corpus or hold-out data testing in this survey are part of this thesis (Weiss and Meurers, 2018; Weiss *et al.*, 2021). One of the limiting factors for cross-corpus evaluations is the lack of comparable reading corpora. Weiss and Meurers (2018) address this issue by compiling two comparable corpora of German media language for adults and children and making them available for research. Similarly, we named making models publicly available to foster ARA research as a primary research desideratum. Accordingly, most readability studies in this thesis make the models accessible either through a web interface (Weiss *et al.*, 2018) or by publishing the models and analysis scripts (Weiss and Meurers, 2022; Weiss *et al.*, 2021) which rely on the publicly available CTAP system which I extended to support analyses of German (see Section 3.3). These papers were also the only studies integrating human processing measures into ARA models. Finally, the ARA approaches included in this thesis were designed to address under-researched target groups including children (Weiss and Meurers, 2018) and L2 readers (Weiss *et al.*, 2021) as well as less commonly researched text units such as sentence-wise readability assessment (see Weiss and Meurers, 2022, in Section 5.3.4).

Chapter 5

Foundational complexity research

5.1 Motivation and core contributions

This chapter focuses on the automatic assessment of written language production and written language reception for inter- and extra-institutional educational contexts. I demonstrate the productivity of the construct of linguistic complexity and its automated, linguistically broad operationalization for predicting a) language proficiency and b) text readability. In the following six corpus-based studies, I investigated the applicability of my integrative approach to broad linguistic complexity modeling for German to APA and ARA for L1 and L2 contexts. I further explored how transferable the approach is to very short language samples.

The remainder of this section focuses on contextualizing the core research questions and findings of each article within the larger frame of this thesis. The other two sections of this chapter present each individual article in more detail, focusing especially on their data, methods, and core findings. Section 5.2 focuses on APA and Section 5.3 on ARA.

5.1.1 Automatic language proficiency assessment

We conducted three studies exploring the value of broad linguistic complexity modeling for automatic L1 and L2 proficiency assessment. Assessing the proficiency of learners through the quality of their written language output has been one of the most prolific application domains of complexity research (see Section 2.2). Yet, manual analyses can only capture a limited range of linguistic dimensions (see Section 3.1.1) and only few automated procedures have been devoted to the systematic study of language competence in German prior to this thesis (see Section 4.2). To address this, we first demonstrated that broad linguistic complexity mod-

eling using automatic feature extraction is indeed beneficial to predict learners' L2 proficiency on the full CEFR scale (A1 to C1/C2; Weiss and Meurers, 2019b). To gain deeper insights into the linguistic characterization of individual proficiency levels and trade-offs between linguistic domains, we further studied for each proficiency level the contribution of individual linguistic domains to the identification of these levels. We found that while lexical and clausal complexity were relevant for predictions across the full CEFR spectrum, discourse and morphological features were notably more informative for the identification of B2 texts than for texts at other levels.

In Weiss and Meurers (2019a), we then turned to the characterization of early L1 academic language development in form of grade level differences from a broad linguistic perspective. For this study, we extended our analysis to consider measures of not only complexity but also accuracy. This allowed us to study developmental trade-offs within complexity domains as well as between complexity and accuracy. We found that in elementary school, pupils developed more in terms of their accuracy. In contrast, in secondary school, the early development of the academic language system shifted towards increasing complexity, especially in the domains of phrasal complexity, lexical complexity, and discourse cohesion. Furthermore, we demonstrated that despite the known impact of task effects on CAF (see Section 2.1.3.2), our linguistically broadly informed models successfully generalized across task prompts with different topics. This is crucial for the reusability of complexity-based classification models.

After having established the value of our approach for longer L1 and L2 writing, we turned to shorter language productions (less than 10 words; Weiss and Meurers, 2021). We successfully predicted L2 proficiency from learners' short answers to reading comprehension questions. We showcased that broad linguistic complexity modeling can be successfully used to predict beginning L2 learners' course levels even from short writing samples. However, we did observe that the cross-task generalizability of models trained on such limited samples was limited, especially when compared to the successful generalization that we observed in Weiss and Meurers (2019a) for models trained on longer data. A second focus of this article was on investigating the robustness of our NLP models and feature extraction algorithms on non-standard language data, in this case produced by beginning learners of German. We could successfully demonstrate that the extraction of complexity measures using our analysis system was hardly impacted by erroneous NLP analyses.

Together, these three studies demonstrate that our integrative approach to linguistic complexity modeling is highly beneficial for the assessment of language proficiency for a broad

range of proficiency levels and language types. Our models have not only achieved high prediction performances. They allowed to zoom in on the linguistic differences between proficiency levels to gain deeper insights into developmental differences across proficiency levels. The studies further showcase that computational linguistic methods can be a valuable instrument to overcome the often reductionist focus of SLA complexity research (see Section 2.1.2). It makes a linguistically broad perspective on complexity feasible while maintaining the required level of robustness even on non-standard language data.

5.1.2 Automatic readability assessment

Turning to the reception of language, we conducted three studies on ARA for German. Even though readability assessment has not been a traditional focus of SLA complexity research, computational linguistic work on ARA has a long tradition of utilizing linguistic complexity features (see Section 2.3). Despite the abundant work on ARA for German, there has been only little work on L2 readers or the assessment of readability below the document level (see Section 4.3.1.2). As our systematic survey further showed, readability formulas have remained the de facto standard for ARA in practice because of the lack of available ARA models leveraging deeper linguistic insights. This has also led to relatively little work discussing the linguistic differences between texts at different readability levels. Yet, these insights would be crucial for deriving recommendations on how to align texts with readers' language skills. To address these limitations, we first set out to demonstrate the value of our broad linguistic complexity modeling approach for ARA. For this, we started with predicting L1 readability in Weiss and Meurers (2018). We used and extended an established corpus to ensure the comparability of our approach to previous work at the time. We showcased that our approach allows to successfully distinguish language targeting adults from language targeting children in educational media language. We also elicited a second corpus of German media language for adults and children to facilitate cross-corpus testing, which we identified as a research desideratum in our background chapter (Section 2.3.3.2) and survey (Section 4.3.1.3). This allowed us to demonstrate the generalizability of our model to a different type of German media language. We further leveraged the insights from our feature-based approach to better understand how language designed for adults and children differed. We found that especially language use, nominal style, and discourse features were important for the distinction of target audiences across media outlets.

After having established the competitiveness of our approach, we focused on providing

more resources for German L2 readers. In Weiss *et al.* (2021), we used our linguistically broad modeling approach to train a highly successful model for multi-level ARA of texts for L2 learners of both German and English. We elicited this corpus specifically for the purpose of addressing the lack of available L2 readability corpora for German as well as the lack of multi-level, multi-lingual readability corpora in general. Moving to a cross-linguistic perspective on L2 ARA, we compared the linguistic differences between reading levels across languages and found several parallels. Specifically, we observed that language use and surface length measures were central for both languages, whereas syntactic complexity was more important for German.

Building on this work, we transitioned from the assessment of longer reading materials to the prediction of L2 sentence readability (Weiss and Meurers, 2022). Besides demonstrating the transferability of our approach to sentence level readability, we specifically focused on comparing our analyses to surface level estimates of readability lacking deeper linguistic insights. This was motivated by the dominance of readability formulas in research using ARA for German (see Section 4.3). We found that broad linguistic complexity modeling outperforms surface-based approaches when predicting the readability of sentences on a continuous scale. For the simpler task of identifying simplified from regular sentences in sentence simplification pairs, however, we saw that also surface feature-based approaches to ARA yielded satisfactory performances. This highlights that for coarse estimations simpler approaches to ARA can indeed suffice but that for precise readability estimates our linguistically informed model yields the best results. Finally, we showcased the use of sentence-wise readability assessment for the analysis of the compositionality of document-level readability, finding evidence that maximum sentence readability rather than average sentence readability is a determining factor of document readability. This is an important insight for work on text adaptation because it showcases that focusing on the most difficult sentences in a text could suffice to decrease overall text readability.

Through these three studies, we have successfully demonstrated that the proposed integrative approach to broad linguistic complexity modeling is a useful tool for the assessment of language reception. Again, the approach allowed to not only predict readability but also to gain deeper insights into the compositionality of text readability, both in terms of the linguistic characteristics of texts at different reading levels and in terms of the link between sentence-level and document-level readability.

5.1.3 Core contributions

With the six studies presented here, we have addressed several known challenges in SLA complexity research. First and foremost, we have demonstrated that the proposed integrative approach to broad linguistic complexity modeling is highly successful for a variety of application contexts linked to language learning. With this, the present work contributes directly to overcoming the often overly reductionist approach to complexity research (see Section 2.1.2). We also utilized this linguistically uniquely broad perspective to make linguistic trade-offs observable in terms of the complexification of different linguistic domains as well as between complexity and accuracy. Taking into account the known task sensitivity of CAF measures, we investigated the robustness of our approach across task contexts in several studies. The findings indicated that the combined evidence from the broad range of linguistic domains yields relatively stable performance estimates provided that enough linguistic material is available to learn generalizable patterns.

From a computational linguistic perspective, these studies have contributed greatly to the two application domains of ARA and APA for German. Across study set-ups, we ensured to address issues of model generalizability to different data sets which not only links back to the concern of task effects but more generally is of uttermost importance for the practical usability of the trained models. To do so, we also addressed the lack of comparable training data for ARA for German. Finally, we have confirmed the robustness of our automatic linguistic analysis on non-standard data, thus addressing the issue of analysis quality on learner or web data. Taken together, these six articles made a substantial contribution to foundational complexity research and computational linguistic work on APA and ARA for German.

5.2 Predicting language proficiency from learner writing

This section summarizes my foundational work on APA for German texts based on broad linguistic complexity modeling. Section 5.2.1 briefly describes all corpora used for the following studies. Section 5.2.2 establishes the value of linguistically broad complexity modeling of German L2 proficiency assessment on longer essays. Section 5.2.3 demonstrates the applicability of the method to identifying differences between young L1 writers in elementary and early secondary school, again on longer essays. Building on these studies, Section 5.2.4 discusses the use of short answers to predict German L2 language proficiency. The sections serve as a concise synthesis of the main results of each study. For a comprehensive descrip-

Table 5.1: *Corpus profile for German Merlin data split by overall proficiency ratings*

	A1	A2	B1	B2	C1	C2
#documents	57	306	331	293	42	4
#sentences	280	2,377	3,834	4,350	597	52
#words	1,956	18,499	39,076	57,262	9,527	984

tion of all study set-ups and detailed reports on the individual study findings, please consult the respective articles listed in Chapter 8.

5.2.1 Corpora and data sets

5.2.1.1 German proficiency corpora for L2 writers

Merlin The Merlin corpus (Wisniewski *et al.*, 2013) is a trilingual cross-sectional learner corpus containing German, Italian, and Czech texts written by L2 learners at beginning to advanced proficiency levels. Throughout this thesis, I will only refer to the German section of the corpus. It consists of 1,033 texts that were elicited during standardized language tests for the five CEFR levels: A1, A2, B1, B2, or C1. Each text was digitized as a diplomatic transcription seeking to faithfully represent the originally hand-written learner text and as a form-based target hypothesis seeking to provide an orthographically and grammatically standardized version with minimal changes. The corpus comes with rich meta information on the learners' background (age, gender, L1s), the task context (test taken, task description), and text quality ratings. All essays were rated on several proficiency scales using the Merlin rating grid (Wisniewski *et al.*, 2013) which was based on the CEFR levels A1 to C2. Each text was rated by two trained raters along several performance dimensions which were combined into a single holistic overall L2 proficiency score. Table 5.1 shows the corpus profile of the Merlin corpus based on the diplomatic transcription.

The corpus was elicited to be balanced across proficiency test levels (each test level is represented by about 200 texts) but not across proficiency ratings. The number of documents across proficiency ratings in Table 5.1 shows how this led to an imbalance of proficiency ratings in the data. The overall proficiency ratings are broadly distributed across the five different test levels. In the most extreme case, a B1 rated text can be an exceptionally well-written text elicited in an A1 test as well as an insufficiently written text elicited in a C1 level test. This introduces a considerable within-proficiency rating heterogeneity into the data, especially be-

Table 5.2: *Corpus profile for CREG-OSU and CREG-KU split by course level*

	CREG-KU				CREG-OSU			
	A1.1	A1.2	A2.1	A2.2	A1.1	A1.2	A2.1	A2.2
#answers	1,995	1,901	1,977	1,966	736	905	809	810
#sentences	2,100	1,982	2,263	2,127	773	1,032	980	1080
#words	7,305	8,384	13,625	15,654	4,323	6,560	7,639	8,967

Table 5.3: *Corpus profile for CREG-7K split by course level*

	A1.1		A1.2		A2.1		A2.2	
	KU	OSU	KU	OSU	KU	OSU	KU	OSU
#answers	742	733	901	905	821	815	814	817
#sentences	794	771	938	1,032	929	988	891	1,089
#words	2,697	4,315	4,128	6,560	5,773	7,748	6,688	9,078

cause at each of the five test levels, texts were elicited by three different task prompts, leading to 15 different task prompts used in the corpus. Ideally, this heterogeneity allows to train classifiers that successfully generalize across very different application contexts. However, it also makes training successful classifiers more challenging.

CREG The Corpus of Reading comprehension Exercises in German (CREG) consists of German L2 short answers to reading comprehension questions elicited in beginning to advanced German courses at two U.S. universities (Ott *et al.*, 2012; Ziai, 2018): The Ohio State University (OSU) and University of Kansas (KU). The task-based corpus contains not only student answers but also the associated reading comprehension questions, reading texts, and teachers' target answers. All learner answers were digitized as a diplomatic transcription. In this thesis, we focused on four subsets of CREG that were elicited in the beginning courses A1.1, A1.2, A2.1, and A2.2. We designed the first three data sets to be balanced across course levels: CREG-OSU consists of 3,259 student answers elicited at OSU. CREG-KU consists of 7,839 student answers elicited at KU. The corpus profiles of these two corpora can be found in Table 5.2. CREG-7K consists of 6,548 student answers elicited in equal parts at OSU and KU. Its corpus profile is shown in Table 5.3. The fourth subcorpus consists of 104 sentences from answers elicited at KU for which Ott and Ziai (2010) created manual dependency annotations using three trained raters. The corpus profile for this data set, which we refer to as CREG-104,

Table 5.4: *Corpus profile for CREG-104 split by course level*

	A1.1	A1.2	A2.1	A2.2
#answers	25	0	32	47
#sentences	25	0	33	52
#words	165	0	221	394

Table 5.5: *Corpus profile for KCT data split by grade levels*

	1st	2nd	3rd	4th	5th	6th	tth	8th
#documents	38	165	276	248	211	251	216	228
#sentences	178	885	2,022	2,616	1,714	2,107	1,513	1,984
#words	1,665	7,725	18,943	26,789	18,825	23,510	18,308	24,366

can be found in Table 5.4. We converted the original dependency annotations to the scheme used by Brants *et al.* (2002) to be able to match them with the output of our NLP pipeline (see Section 3.2). We also manually augmented the data set with reference annotations for lemmas and morphological inflection (case, number, gender, person, tense, verb mode, degree of comparison), see Weiss and Meurers (2021, Section 4.3.1) for details.

5.2.1.2 German proficiency corpora for L1 writers

Karlsruhe Children's Texts The KCT corpus by Lavalley *et al.* (2015) is a cross-sectional corpus of German texts written by pupils who attended German elementary or secondary school at the time of writing. It consists of 1,701 texts elicited in first to eighth grade at elementary school and two types of secondary school: German *Hauptschule* which is the basic vocational secondary school track and German *Realschule* which is the advanced vocational secondary school track. All texts were elicited with one of two age-appropriate task prompts using different topics. At elementary school, pupils were asked to continue one of two narratives: one about playing in a park or one about a wolf learning how to read. At secondary school, pupils had to write a fictional text about a day spent with their idol or their life in 20 years. The corpus contains faithful transcriptions of students' texts as well as a form-based normalization. The corpus additionally includes rich annotations for error types regarding word segmentation, word choice, grammar, and legibility. Table 5.5 shows the corpus profile of the KCT corpus split by grade levels. As can be seen, the number of texts varies greatly between grade levels. To obtain a more homogeneous distribution, we always grouped two

adjacent grade levels (1st/2nd, 3rd/4th, 5th/6th, 7th/8th) for our analyses in Weiss and Meurers (2019a).

5.2.2 Broad linguistic modeling for German L2 proficiency assessment

In our first study on APA for German (Weiss and Meurers, 2019b), we focused on demonstrating how broad linguistic complexity modeling provides added value to the task of L2 proficiency classification over approaches informed by linguistically homogeneous feature sets. As a second research focus, we compared the informativeness of individual feature domains for the distinction of specific proficiency levels on the full CEFR range from beginning to advanced learners. This allowed us to better understand at what point in the developmental process certain linguistic domains evolve. Traditional SLA complexity research has successfully shown that throughout the developmental trajectory of learners, linguistic domains evolve at different speeds and time points (see Section 2.1.3.1). These studies have mostly focused on comparing relatively few features from the domains of syntax, lexicon, and language use. Our broad linguistic complexity modeling approach in contrast offered the possibility to compare a variety of linguistic domains simultaneously on the full developmental scale from beginning to advanced L2 learners of German.

As data basis for our analysis, we used the diplomatic transcriptions in the Merlin corpus (see Section 5.2.1.1). The unique size and breadth of the corpus allowed us to train a proficiency classifier for beginning to advanced L2 learners of German. We used the holistic overall CEFR ratings as training labels to distinguish five proficiency levels: A1, A2, B1, B2, and C1/C2.[1] We extracted a total of 400 complexity features covering the linguistic domains of syntax (separated into phrasal and clausal complexity), lexicon, morphology, discourse, language use, and human language processing. We calculated these features on the Merlin data using the original monolingual complexity analysis system introduced in Section 3.2. We then ranked features by their informativeness for the distinction of proficiency levels using information gain. This allowed us to form a total of twelve different feature sets: one per linguistic domain ($N = 7$), one combining all features, and four using the 200, 150, 100, and 50 most informative features from our previous ranking. With each of these feature sets, we trained one classifier using 10-CV. All classifiers used the SMO algorithm (Platt, 1998) which is known for its stability when used with highly inter-correlated features. We compared their

[1] Since the Merlin corpus only contains four essays receiving an overall rating of C2, we did not distinguish between C1 and C2 ratings.

Table 5.6: *Performance of L2 proficiency models trained in Weiss and Meurers (2019b) in terms of overall accuracy and level-wise F1 score (highest performance each comparison marked with bold font)*

Metric	Level	Maj	All	150	Lex	Cla	Phr	Mor	Use	HLP	Dis
Accuracy	A1–C1/2	32.0	68.1	**70.0**	67.6	63.8	62.1	59.7	59.3	53.7	64.7
F1 score	A1	0.0	40.8	**45.7**	27.3	12.3	3.3	0.0	3.4	0.0	6.7
	A2	0.0	70.2	**73.9**	73.2	68.9	68.3	63.1	63.6	60.1	69.5
	B1	48.5	65.8	**67.1**	62.1	58.6	56.4	54.0	54.4	49.1	58.2
	B2	0.0	74.4	**77.4**	75.9	73.7	71.7	72.0	69.5	61.1	75.1
	C1/2	0.0	**31.2**	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0

performances with each other and with a naive baseline (always predicting the majority class: B1) in terms of their overall accuracy as well as in terms of their proficiency level-wise precision, recall, and F1 score. All analyses were carried out using the WEKA machine learning toolkit (Hall *et al.*, 2009) and the statistical programming language R (R Core Team, 2022) using the packages `ggplot2` (Wickham, 2016) and `gridExtra` (Auguie, 2022).

Classifier evaluation

Table 5.6 summarizes the classification performance of all models in terms of their overall accuracy and their proficiency level-wise F1 scores. Instead of all four models using the information gain ranking feature selection, only the best-performing one using the 150 most informative features was reported here. We see that all models outperformed the majority baseline (Maj.) with an accuracy of 32.0%. The classifier using the 150 most informative features (150) achieved an overall accuracy of 70.0% and outperformed the other classifiers. When inspecting which linguistic domains were being represented in this data-driven feature set, we saw that it was informed by features from all linguistic domains. This showcases that the full range of our broad linguistic modeling approach contributed to the higher performance of the classifier compared to the ones using homogeneous feature sets. It is noteworthy that the performance of the most successful homogeneous classifier—which used 38 lexical complexity features (Lex)—was 67.6%, i.e. only 2.4% below the best-performing model despite using a much smaller and easier to compute feature set. Yet, the proficiency level-wise performance evaluation revealed that the diverse model was indeed superior in distinguishing proficiency levels from beginning to advanced when compared to any of the other models including the

lexical model.[2] The lexical model performed much worse for the identification of A1 rated essays and failed for C1/2 rated essays. This was a general pattern that we observed across all models informed by homogeneous feature sets. When summarizing this difference in terms of the weighted average F1 score, the impact of this performance difference between the model using the 150 most informative features ($F1_M = 68.1\%$) and the lexical model ($F1_M = 64.6\%$) also became apparent on the macro level. The evidence clearly showed that a broad linguistic complexity modeling approach is highly beneficial for L2 proficiency assessment on the full CEFR scale.

Our second research question concerned performance differences within the homogeneous proficiency models across proficiency levels to gain a better understanding of which linguistic domains are suited to characterize learners at the respective proficiency levels. We found that lexical and clausal complexity systematically performed best among the homogeneous models across CEFR levels. Also discourse features were more successful than the other linguistic domains, especially for B2 essays. The other homogeneous feature domains showed much less discriminatory power across proficiency levels, except for the morphological model performing remarkably well for B2 rated essays. Interestingly, all homogeneous models failed to correctly identify advanced learners (C1/2). They also performed poorly for beginners' essays (A1). These two classes were also the most challenging to distinguish due to their under-representation in the Merlin corpus, yet, the linguistically diverse models handled these cases much better than the homogeneous models despite training with the same distribution of classes.

5.2.3 Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school

In Weiss and Meurers (2019a), we used our broad linguistic modeling approach to track the early development of German L1 academic language. Unlike in our previous studies on L2 proficiency, we measured both complexity and accuracy to characterize pupils' development. This allowed us to investigate potential developmental trade-offs between complexity and accuracy. We focused on the first eight years of schooling in the German education system.

[2]Only the model using all features (All) outperformed the model using the 150 most informative levels in one instance, namely for the identification of learners at the C1/C2 level. Since the performance difference between both models in terms of their overall accuracy was relatively small, one might in fact prefer the model using all features over the one using only the 150 most informative features.

This included writings from elementary school (first to fourth grade) and from the first four years of two types of secondary school (see below and Section 5.2.1.1 for details). Our studies focused primarily on grade level differences but also payed special attention to the transition between elementary and secondary school. This step is known to be a formative turning point in contrast to the more gradual transition between grade levels within school types (e.g., Riebling, 2013). Most research on ALPS for German has focused on adults' L2 development or their L1 academic language development at university (see Section 4.2). Less work has been dedicated to early stages of academic language development that take place in the first years of schooling. Existing studies on L1 writing development mostly focused on the development of students' accuracy (Göpferich and Neumann, 2016; Laarmann-Quante *et al.*, 2019; Lavalley *et al.*, 2015). In this article, we addressed the lack of research on early academic language development in terms of complexity with three main research contributions. First, we built successful classification models for the distinction of texts written by pupils at different grade levels. Second, we used the linguistic insights that we obtained from our feature-based machine learning approach to describe the developments in pupils' writing from a broad linguistic perspective. Third, taking into account the influence of task effects on CAF (see Section 2.1.3.2), we demonstrated the cross-topic generalization of our models to unseen writing topics.

We used a subset of the KCT corpus (see Section 5.2.1.2) as empirical basis for our analyses. The error annotations in the KCT corpus allowed us to consider accuracy as a second dimension of language performance. We wrote a script that inferred 37 error rate measures from the existing annotations in the KCT corpus.[3] The error rate features covered measures of incorrect word choice, spelling mistakes, and punctuation errors. We also extracted 308 features of linguistic complexity using our original complexity analysis system (see Section 3.2). This included features from the linguistic domains of syntax, lexicon, and morphology, features of discourse cohesion, and psycho-linguistic features of language use and human processing. We extracted complexity measures from the normalized transcription of pupils' writings to better delineate complexity and accuracy effects. We removed all features exhibiting near zero variance in the data and computed the z-scores of the remaining features. From these, we formed eleven different feature sets: one for each complexity domain (separating phrasal and clausal complexity), one for the error rate features, one combining all complexity features, one combining all complexity and accuracy features, and one adding also meta information

[3]The script is available at `https://github.com/zweiss/KCTErrorExtractor`.

on topic and school type (if applicable, see below). Using these feature sets, we trained classifiers to predict grade levels in three different classification tasks: i) a four-way classification of elementary and secondary school writing (1st/2nd, 3rd/4th, 5th/6th, 7th/8th), ii) a binary classification of elementary school writing (1st/2nd, 3rd/4th), and iii) a binary classification of secondary school writing (5th/6th, 7th/8th). We trained the two binary classifiers to examine the differences between pupils' development in elementary and secondary school. All classifiers used the SMO algorithm (Platt, 1998) and were trained and tested with ten iterations of 10-CV. Both binary classifiers were additionally trained on one topic prompt and tested on the other to test the cross-topic generalization of the models. We compared the performance of all models in terms of accuracy against two baselines: a majority baseline and a surface feature classifier using only text length and word length features. Finally, we ranked features based on their information gain for the distinction of grade levels separately on the elementary school data and on the secondary school data. We compared the most informative features from each feature domain for both rankings to gain an understanding of the linguistic differences between grade levels and school types. All analyses were carried out using the WEKA machine learning toolkit (Hall *et al.*, 2009) and the statistical programming language R (R Core Team, 2022) using `tidyverse` (Wickham *et al.*, 2019).

Classifier evaluation

For the four-way classification, all feature-based models clearly outperformed the two baselines which achieved an accuracy of 32.1% each. Our best-performing model used all features plus meta information on school type and writing topic. It obtained an accuracy of 72.7%. Also the model without meta information achieved an accuracy of 71.0%. The combination of complexity measures and accuracy measures had a greater impact on model performance: The complexity model reached only an accuracy of 68.4%. The models using only features from individual feature domains reached a systematically lower accuracy ranging from 42.2% (human processing) to 61.3% (phrasal complexity). We found that the phrasal, lexical, and discourse models and (with some distance) the morphological model all achieved higher accuracies than the error rate model. The models using psycho-linguistic features (human processing and language use) performed worst. This suggests that these measures play only a minor role in early L1 academic language development.

The results for the binary classifier trained on the secondary school writings were similar in the sense that again all models significantly outperformed the baseline models except for the

human processing model. Again, the model combining all features with meta information on task prompt and school type performed best with an accuracy of 65.7% against the baseline models performing both at around 51%. Adding meta information accounted for a boost in accuracy of 2.5%. However, combining complexity and error rate features did not improve performance for this subset of the data. In fact, when comparing the relevance of the individual linguistic domains for the distinction of grade levels in secondary school, we saw that all complexity models outperformed the error rate model except the morphological model (performing at the same level) and the human processing model (which was uninformative). This indicates that accuracy develops less systematically across grade levels in secondary school than complexity.

This stands in sharp contrast to our observations for elementary school writing. Again, the combination of all features and meta information on task prompt[4] performed best with an accuracy of 82.8% compared to the baseline models achieving an accuracy of 71.7% each. Adding the information, however, did not significantly improve performance (82.6%) and the error rate model alone achieved an accuracy of 81.6%, which was a significant but small performance difference. We also observed that only the phrasal, discourse, and lexical model significantly outperformed the baseline models whereas the morphological, language use, and human processing models did not. This shows that elementary school writing develops predominantly in terms of accuracy, even though there also seems to be some development in the domains of lexicon, phrasal complexity, and discourse.

Cross-topic evaluation

In a final machine learning experiment, we tested the cross-topic generalization of our models. Because writing topics in KCT do not overlap between elementary and secondary school, we used the binary models for our evaluation. We compared the performances of the complexity model, the error rate model, and the model combining all features with and without meta information. The results showed that models on both data sets generalized well replicating the performance differences that we observed in the previous evaluation. On the elementary school data, the accuracy model not only outperformed the other two models but also performed at the same level as in the previous classification experiment (81.6%). Also the complexity and the combined model performed well in the cross-topic experiment, despite a drop in accuracy of up to ten percent compared to the model trained on both topics.

[4]Within elementary school, there are not different school types making this information superfluous.

As for secondary school, the model combining all features and the meta information on school type performed best again, with a drop in performance of only three percent compared to the model trained on both topics. The model without meta information and the complexity model, too, remained relatively stable with drops of performance around only four percent. The error rate model again showed no drop in performance compared to the previous experiment but also performed worst with an accuracy of only 55.2% compared to the majority baseline of 50.0%. In summary, this shows that all complexity models combining evidence from a broad range of linguistic domains generalized reasonably well across topics. Accuracy measures proved to be fully stable but of limited importance for secondary school academic language development from fifth to eighth grade.

Zooming in on individual features

Finally, we zoomed in on the most informative features of each linguistic domain and compared them across grade levels in both school types to better understand their developmental trajectory. Even though not all patterns observed through our comparison were clearly interpretable, we saw mostly examples of systematic development in the data. Accuracy systematically increased from 1st/2nd grade to 7th/8th grade. Similarly, lexical diversity showed a systematic development in terms of Yule's k and the coverage of noun phrase modifiers increased significantly for both school types. In line with our previous observations, other features developed exclusively in secondary school. This held specifically for the use of derived nouns, the number of conjunctional clauses per sentence and verbs per t-unit, and the use of vocabulary associated with newspaper writing, all of which increased in secondary school. Interestingly, also human processing measures showed a small but significant increase across grade levels, even though they contributed little in the classification experiments.

5.2.4 Analyzing the linguistic complexity of German L2 short answers

In Weiss and Meurers (2021), we shifted our analysis of German non-standard L2 writing for proficiency assessment from longer to very short texts. Previous research had predominantly focused on longer texts types such as essays, letters, narratives, and descriptions (see Sections 2.2.3 and 4.2.1.2). Yet, being able to place L2 learners on a proficiency scale with short writing samples would be very useful in practice because shorter text samples are easier to obtain. Also, given the known influence of task factors on CAF (see Section 2.1.3.2),

it is important to study learner language across a range of writing tasks. Thus, we analyzed L2 learners' short answers (less than 10 words) to different reading comprehension tasks and studied the models' cross-task generalization. The second core objective in this paper was to quantify the impact of non-standard language on our NLP annotations ranging from POS tagging to dependency parsing and the subsequent calculation of complexity measures based on these annotations.

We based our analyses on the CREG corpus (see Section 5.2.1.1). To obtain L2 proficiency classifiers and test their generalization across task contexts, we extracted 297 complexity features from the student answers in CREG-KU, CREG-OSU, and CREG-7K. We calculated all features using our original complexity analysis system (see Section 3.2). After removing insufficiently variable features, we obtained a final set of 147 features covering all linguistic domains for which we calculated z-scores. Each data set was split into a training set (70%), a development set (20%), and a test set (10%).[5] To test the generalization of models to held-out data, we tested the KU model on CREG-OSU and vice versa. We conducted two additional machine learning experiments in which we used a 70/10/10/10 split of training set, development set, regular test set, and held-out test set. The held-out test set either consisted of held-out reading comprehension questions or of held-out reading texts (including their corresponding reading comprehension questions). These additional experiments allowed us to evaluate the generalization of our proficiency models for increasing degrees of held-out data. We trained an Ordinal Random Forest (ORF) on each of the training data sets to predict the course level. All models were evaluated on their respective test sets (regular, held-out questions, held-out reading texts, cross-university). Model performance was always compared to the majority baseline on the respective test sets and quantified in terms of overall accuracy.

To quantify the effect of non-standard language on the quality of NLP annotations and the subsequent calculation of our complexity features, we followed a three-step procedure: First, we applied our NLP pipeline (see Section 3.2) on CREG-104 and evaluated the performance of automatic POS tagging, lemmatization, morphological analyses, and dependency parsing using the data set's manual annotations as reference. This allowed us to understand the general performance of the NLP tools on the data. Second, we extracted complexity features based on these annotation layers using our original analysis system ($N = 93$) utilizing a) the manual reference annotations and b) the automatic NLP annotations. Of these, 69 were variable across the 104 sentences. We calculated the z-scores for these features and compared the

[5] We chose a train-test split over 10-CV because of the substantial size of the three CREG data sets.

root mean squared difference (RMSD) between features based on these two types of annotations. This allowed us to understand the impact of errors in the automatic annotations on the calculation of complexity features. Third, we trained a classifier using only the 69 variable features on CREG-KU with a 70/20/10 split and compared its performance on the test set with the performance on CREG-104 with features based on the automatic or the manual annotations. This allowed us to quantify the impact of errors in the automatic annotations on our prediction of course levels. For all analyses, we used the statistical programming language R (R Core Team, 2022) and RStudio (RStudio Team, 2022). We used the packages `tidyverse` (Wickham *et al.*, 2019), `caret` (Kuhn, 2022), `kernlab` (Karatzoglou *et al.*, 2022), `e1071` (Meyer *et al.*, 2022), `ranger` (Wright and Ziegler, 2017), `randomForest` (Meyer *et al.*, 2022), `PerformanceAnaytics` (Peterson and Carl, 2020), `caretEnsemble` (Deane-Mayer and Knowles, 2019), and `doParallel` (Microsoft Corporation and Weston, 2022).

Classifier performance and generalization

All three models from the 70/20/10 split achieved high classification accuracies on their respective test sets that clearly outperformed their respective majority baselines (KU = 84.6% against a baseline of 25.9%, OSU = 80.6% against a baseline of 27.7%, 7K = 74.9% against a baseline of 26.4%). However, neither the OSU model nor the KU model generalized to the respective other corpus. Both showed accuracies identical or close to the respective baseline models. Yet, the success of the 7K model demonstrated that it is possible to learn common course level characteristics between both universities. We investigated this discrepancy further by inspecting the performances of the models on the held-out questions and held-out reading texts test sets.[6] We saw that models generalized to both held-out questions and held-out texts but that their accuracy systematically declined with increasing dissimilarity to the training data. On the held-out questions test set, the KU classifier reached an accuracy of 57.3%, the OSU classifier an accuracy of 61.8%, and the 7K classifier an accuracy of 54.4%. On the held-out texts test set, the KU classifier reached an accuracy of 40.5%, the OSU classifier an accuracy of 40.7%, and the 7K classifier an accuracy of 40.1%. Taken together, these findings show that our classifiers generalized to some extent. Yet, the generalization that can be obtained for such short texts seems to be somewhat limited. To use such classifiers in practice,

[6]The performance of the models trained with a 70/10/10/10 split on their respective regular test sets and the cross-university test sets was equivalent to their performance in the 70/20/10 split (as to be expected). They were not reported here. I focused on their performance on the held-out questions and held-out texts sets.

it would therefore be ideal to train classifiers on answers responding to the reading tasks that they are supposed to be used for later on. This stands in contrast to the cross-topic robustness for the classification of longer texts in Weiss and Meurers (2019a).

Robustness of complexity modeling on non-standard language

Overall, only 20.4% of automatically annotated sentences fully matched the manual reference annotations. This means that most sentences contained at least partially incorrect analyses. Yet, the individual NLP components in our pipeline performed reasonably well on CREG-104. The POS tagging accuracy ranged between 92.7% and 99.0% with non-finite verbs and adjectives receiving the lowest scores. Regarding lemmatization, we mostly observed accuracies above 90% except for adjectives (84.0%) and finite verbs (85.7%). The accuracy of morphological analyses ranged from 75.7% to 100%. We found the morphological analyses of nouns, adjectives, and non-finite verbs to be more accurate than the analyses of non-finite verbs. In the ensemble NLP system that we used (see Section 3.2), there was a direct link between the lower performance of the morphological analyzer for non-finite verbs and the previously mentioned difficulties that the POS tagger had in correctly labeling non-finite verbs which were often labeled as finite verbs. For adjectives and nouns, we found the labeling of case and gender to be most challenging. Finally, for the dependency analyses we observed reasonably good unlabeled attachment scores and fair labeled attachment scores except for separable verb particles (25.0%) and relative clauses (50.0%). We found that object relations were labeled less reliably than subject relations, which could only be partially explained by errors in the morphological analysis of case. To summarize, these findings show that the automatic analysis of non-standard data is possible but challenging for our NLP pipeline.

In the second step, we focused on the impact this had on the calculation of complexity measures. We compared the RMSD between z-scores of features calculated using the manual reference annotations versus the automatic annotations. This allowed us to quantify the difference introduced by using the different annotation bases in terms of standard deviations. Overall 81.2% of features showed no difference (10/69) or a weak difference (RMSD ≤ 0.5; 46/69). A total of eleven features showed a medium difference ($0.5 <$ RMSD ≤ 1). These were predominantly based on the assignment of subject or object dependency relation or case labels. Two features entirely based on the correct assignment of subject and object dependency relation labels showed a substantial deviation ($1 <$ RMSD ≤ 2). We did not find any extreme deviations (RMSD > 2). These findings allowed us to reach two conclusions: First,

the previously noted issues with labeling objects and subjects during dependency parsing and case assignment for nouns and adjectives in fact impacted the calculation of complexity measures. Second, however, it also demonstrates that the other challenges noted in the previous assessment hardly influenced the calculation of features. In particular, this put into perspective our earlier finding that nearly 80% of sentences contained at least one annotation error.

Whether or not the weak to substantial differences in the feature calculation are acceptable, depends on the purpose for which complexity measures are being calculated. In Weiss and Meurers (2021) and throughout this thesis, we used complexity measures predominantly as features for training machine learning classifiers. Thus, we tested the impact of using manual or automatic annotations for feature calculation on the prediction performance of a model trained on CREG-KU. Our results showed that the performance of the model is comparable between the held-out KU test set, CREG-104 with automatic annotations, and CREG-104 with reference annotations. Thus, the observed differences between complexity measures based on automatic and manual annotations did not seem relevant for the purpose of training linguistically broadly informed classifiers. That being said, our findings were somewhat limited a) by the small size of the CREG-104 data set and b) because the relative simplicity of the sentences made the syntactic analysis easier and limited the range of linguistic forms that we could observe. Also, we did not have reference annotations for constituency parses which prevented us from assessing the robustness for all features. It remains to be empirically tested to what extent our results can be generalized to more complex sentences and features based on constituency parsing. To our knowledge, currently, there is no suitably annotated corpus available for this. Despite these limitations, our evaluation covered a broad range of NLP tasks and linguistic domains which we evaluated on authentic non-standard data. We are therefore confident that our findings have a certain validity beyond the context of the CREG data.

5.3 Identifying competence-adaptive text input for learners

This section summarizes my foundational work on ARA for German using broad linguistic complexity modeling. Section 5.3.1 briefly describes all readability corpora used for my research on ARA. Unlike in Section 5.2, most corpora used in this section were elicited by us specifically to answer our research questions and to address the shortage of German ARA corpora. Section 5.3.2 focuses on ARA of German media language for L1 readers and Section 5.3.3 on ARA for German and English L2 texts. Section 5.3.4 moves the complexity-

based ARA approach from the document-level to the sentence-level. This entire section focuses on the main results from each paper and on linking the collected findings into a coherent picture. For a comprehensive description of all study set-ups and detailed reports on the individual study findings, please consult the respective articles in Chapter 8.

5.3.1 Corpora and data sets

5.3.1.1 German readability corpora for L1 readers

GEO/GEOlino The GEO/GEOlino corpus (Weiss and Meurers, 2018) is a binary, leveled corpus of German media language. It consists of articles from the German educational monthly magazine GEO targeting adults and its adaptation for children (6 to 14 years) GEOlino. GEOlino articles are not simplifications of corresponding GEO articles. Instead, GEO and GEOlino are two independent magazines by the same publisher with similar but neither identical nor coordinated contents targeting two different audiences: adults and children. This makes these articles a very valuable source of data for German ARA. Hancke *et al.* (2012) compiled a first version of this corpus consisting of 4.603 articles by crawling the sites `www.geo.de` and `www.geolino.de`. We followed their set-up and crawled an updated and nearly twice as large version of the corpus. It contains overall 8,263 articles on the topics crafting, humanities, nature, reviews, technology, and travel after clean-up and removal of texts with less than 15 words. As the original corpus, the full new GEO/GEOlino corpus is not balanced between GEO ($N = 4,999$) and GEOlino ($N = 3,264$) texts. To account for this, we created the balanced data set GEO/GEOlino$_S$. It consists of 2,480 texts on topics that were represented in both, GEO and GEOlino: humanities, nature, and reviews. Table 5.7 contains the corpus profile of GEO/GEOlino$_S$, including the total number of documents and the median number of sentences and words per document for GEO$_S$ and GEOlino$_S$ articles.

Tagesschau/Logo corpus The Tagesschau/Logo corpus (Weiss and Meurers, 2018) is a binary, leveled corpus of German media language for information and education. It consists of subtitles from two major German daily news broadcasts that aired from December 2015 to January 2017: the *Tagesschau* by the German public-service television network ARD and *Logo!*, a news service for children (age 6 to 14) provided by the German public-service television network ZDF. For all broadcasts, the subtitles were cleaned from meta-comments on non-verbal audio cues for hearing impaired audiences. While multiple editions of *Tagesschau*

Table 5.7: *Corpus profiles for the L1 readability corpora used in this thesis: GEO/GEOlino$_S$, Tagesschau/Logo, GEO/GEOlino$_4$, and Tagesschau/Logo$_{\frac{1}{5}}$*

	GEO$_S$	GEOlino$_S$	Tagesschau	Logo
#documents (total)	2,480	2,480	421	415
#sentences (median)	23	25	167	125
#words (median)	383	350	1,631	1,322
	GEO$_4$	GEOlino$_4$	Tagesschau$_{\frac{1}{5}}$	Logo$_{\frac{1}{5}}$
#documents (total)	420	420	2,049	2,049
#sentences (median)	112.5	122.5	32	24
#words (median)	1,797	1,741	325	259

air throughout the day, *Logo!* is broadcasted once per day. To obtain a comparable number of documents, the Tagesschau/Logo corpus only includes the main edition of *Tagesschau* that airs every evening at 8pm. Tabe 5.7 contains the corpus profile of Tagesschau/Logo. The minor mismatch in the number of documents from *Tagesschau* and *Logo!* is due to the fact that unlike *Tagesschau*, *Logo!* does not air on certain holidays.

Aligning GEO/GEOlino and Tagesscha/Logo The GEO/GEOlino$_S$ corpus and the Tagesschau/Logo corpus (both Weiss and Meurers, 2018) were published together to facilitate cross-corpus comparisons of German ARA models. While both corpora target audiences at similar age ranges and both contain German media language for information dissemination, they differ considerably in their corpus profiles. This becomes apparent when comparing their profiles in Table 5.7. The GEO/GEOlino$_S$ corpus contains more than five times as many documents than the Tagesschau/Logo corpus. Yet, individual texts in GEO/GEOlino$_S$ are about four times shorter in terms of the median number of sentences and words than texts in the Tagesschau/Logo corpus. To compensate for these differences, we created two modified data sets. GEO/GEOlino$_4$ reduces the number of articles to 840 while simultaneously lengthening individual texts by appending up to four GEO/GEOlino$_S$ texts from the same topic domains and sampling 420 appended GEO texts and 420 appended GEOlino texts. Tagesschau/Logo$_{\frac{1}{5}}$ increases the number of subtitles while simultaneously shortening the length of individual texts by splitting each original Tagesschau and Logo transcript into five equi-sized partitions and sampling 2,049 partitions from Tagesschau and 2,049 partitions from Logo. Table 5.7 contains the corpus profile of both corpora. As can be seen, the modification of both corpora made the

Table 5.8: *Corpus profiles for Spotlight-DE and Spotlight-EN*

	Spotlight-EN			Spotlight-DE		
	Easy	Medium	Advanced	Easy	Medium	Advanced
#documents (total)	1.030	1.528	1.030	763	509	174
#words per document						
Mean	206	588	606	236	665	892
Standard deviation	166	555	509	235	769	537
Median	137	493	489	137	448	524
Minimum	53	23	26	60	72	91
Maximum	877	4.497	2.940	1.469	5.605	4.161

GEO/GEOlino$_4$ and the Tagesschau/Logo corpus comparable in terms of their profiles. The same holds for Tagesschau/Logo$_{\frac{1}{5}}$ and GEO/GEOlino$_S$.

5.3.1.2 German readability corpora for L2 readers

Spotlight corpora The Spotlight corpus (Weiss *et al.*, 2021) consists of 3.285 English articles (Spotlight-EN sub corpus) and 1.446 German articles (Spotlight-DE sub corpus) at three difficulty levels (easy, medium, advanced). All articles come from monthly language learning magazines published by the Spotlight publisher. Data from the publishers' Italian, Spanish, and French L2 magazines is also available and currently being prepared to extend the corpus. The publisher targets learners at specific CEFR levels. According to Spotlight, easy texts are designed for learners at the A2 level, medium texts for learners at the B1 and B2 level, and advanced texts for learners at the C1 level. However, this link has not yet been independently verified in an empirical study. Table 5.8 summarizes the corpus profile for Spotlight-EN and Spotlight-DE sub corpora.

TextComplexityDE The TextComplexityDE corpus by Naderi *et al.* (2019b) consists of 1,119 sentences from 23 Wikipedia and articles and two articles in German *Leichte Sprache* (engl. "easy language"). It is the only corpus in this section that was not elicited as part of this thesis. In a large scale annotation experiment, all sentences were rated by 267 German L2 learners for three dimensions on a 7-point Likert scale: readability, lexical difficulty, and understandability. To obtain one readability estimate per sentence, Naderi *et al.* (2019b) aggregated all human ratings for a sentence into a single mean opinion score through averaging. The resulting mean average opinion readability score – which is the score used in this thesis

Table 5.9: *Corpus profiles for TextComplexityDE corpus*

	Mean	Std.	Min.	Max.
Readability score	3.02	1.18	1.00	6.33
Words / sent.	20.08	10.62	4.00	63.00
Syll. / word	2.07	0.35	0.96	4.00

in Section 5.3.4 (Weiss and Meurers, 2022) – ranges from 1 to 6.33. Table 5.9 contains the corpus profile for the TextComplexityDE corpus.

The corpus also contains a sub corpus of 249 sentence pairs of simplified sentences and their original in regular German. Simplifications were obtained through human subjects who also indicated whether or not their changes weakly or strongly simplified the original sentences. Simplified sentences are not included in the corpus profile in Table 5.9.

5.3.2 Modeling the readability of German targeting adults and children

As an initial approach to broad linguistic complexity modeling for German ARA, we studied the binary distinction of media language for German L1 speaking adults/adolescents (older than 14 years) and children (6 to 14 years) in Weiss and Meurers (2018). The goal of this paper was two-fold: First, it investigated which (if any) features were informative for the distinction of German media language targeting adults/adolescents from language targeting children on both data sets. This allowed us to understand if and how publishers adjust the language in their materials systematically to different target audiences—something schoolbook publishes have been shown to not always accomplish on a broad linguistic scale (Berendes *et al.*, 2018). Second, it focused on building a state-of-the-art binary classifier for German media language that generalizes across corpora. Cross-corpus performance evaluations are central to ensure that a model is applicable to new data, thus corroborating the practical relevance of the model.

Cross-corpus evaluation studies require the availability of at least two comparable reference corpora, which were not available for German prior to this work. We compiled two corpora of German media language to address this shortage (see Section 5.3.1.1): The GEO/GEOlino corpus reproduces and enlarges the original GEO/GEOlino corpus compiled by Hancke *et al.* (2012) and consists of magazine articles written for adults and children. The Tagesschau/Logo corpus consists of news broadcast subtitles for adults and children and was newly compiled for this study. Combining these two corpora allowed us for the first time to control for the cross-

corpus generalizability of a German ARA model. On this data, we extracted 400 features of linguistic complexity for each article from all linguistic domains introduced in Section 2.1.2. They were calculated using the original complexity analysis system introduced in Section 3.2 and used throughout the two studies that are part of this paper: One focusing on he identification of the most informative features on both data sets and the one on the training and cross-corpus evaluation of two feature-based machine learning classifiers. Both classifiers used the SMO algorithm (Platt, 1998). The models were trained using the Weka machine learning toolkit (Hall *et al.*, 2009). We also used Weka to calculate information gain using the information gain attribute evaluation algorithm with a ranking search. For all further analyses, we used the statistical programming language R (R Core Team, 2022) and RStudio (RStudio Team, 2022).

Feature informativeness

The information gain ranking revealed that 79.0% (316/400) of features were informative for the distinction of target audiences on the GEO/GEOlino$_S$ data and 88.3% (353/400) on the Tagesschau/Logo data. When inspecting the 20 most informative, not highly inter-correlated features for each data set, we observed a notable difference in the range of average feature merits. On Tagesschau/Logo the merit was considerably higher (ranging from $.50 - .98$) than on GEO/GEOlino$_S$ (ranging from $.11 - .33$) – so much so that the highest merit in the top 20 selection for GEO/GEOlino$_S$ in fact fell below the lowest merit in the top 20 selection for Tagesschau/Logo. This is particularly remarkable because the selection also spanned a much wider range of original rankings for Tagesschau/Logo than for GEO/GEOlino$_S$. As for the types of linguistic features being identified as informative, all feature domains were represented at least once in the top 20 feature selection on both data sets except for human language processing. This shows that texts for different target audiences differ on both data sets in terms of a broad range of features. Overall 55% of features in the top 20 feature selection came from the domains of language use and discourse. Also features tied to noun use and nominal complexity were ranking high on both data sets. All in all, the broad linguistic adaptation of texts to their target audience on both data sets showed clear parallels despite some corpus-specific differences. The findings show that different German media language publishers differentiate the linguistic design of their materials based on their target audiences in comparable ways, going well beyond surface text characteristics such as text and word length.

5.3.2.1 Classifier and cross-corpus evaluation

The SMO classifier trained and tested on GEO/GEOlino$_S$ with 10-CV using all 400 features obtained an average accuracy of 89.5% across folds ($SD = .09$), which was well above the random baseline (50.0%). The classifier trained and tested with 10-CV on Tagesschau/Logo obtained an average accuracy of 99.9% across folds ($SD = .04$), again outperforming the random baseline (50.0%). Although the performance of the Tagesschau/Logo classifier was near perfect in 10-CV and the low standard deviation across folds did not seem to indicate any overfitting, the classifier did not generalize to the GEO/GEOlino$_S$ data in the cross-corpus evaluation: It achieved an accuracy of only 52.2%. When contrasting the models' predictions with the gold standard labels, we saw that the model predicted 97.5% of all GEO/GEOlino$_S$ texts to be targeted at children, meaning that the model underestimated the difficulty of GEO texts. In contrast, the GEO/GEOlino$_S$ classifier achieved an accuracy of 98.9% on Tagesschau/Logo in the cross-corpus evaluation. This was not only much higher than its performance during 10-CV on GEO/GEOlino$_S$, it was also close to the performance of the Tagesschau/Logo classifier during 10-CV. This showed that the GEO/GEOlino$_S$ classifier generalized exceptionally well to the unseen data set of German media language. The lack of generalizability for the Tagesschau/Logo classifier, therefore, does not seem to have been caused by a lack of common learnable linguistic differences between German media language targeting adults and children.

Another potential cause for the performance difference is that GEO/GEOlino$_S$ might be a better training corpus due to its considerably larger number of training documents. After all, Tagesschau/Logo has not even a fifth of the size of GEO/GEOlino$_S$ in terms of the number of documents. At the same time, the median number of words and sentences per text are approximately four times larger in Tagesschau/Logo than in GEO/GEOlino$_S$. To ensure that neither of these two differences in the corpus profiles caused the difference in model performance, we conducted a second classification experiment using the GEO/GEOlino$_4$ corpus and the Tagesschau/Logo$_{\frac{1}{5}}$ corpus. The corpus profile of GEO/GEOlino$_4$ was aligned with the Tagesschau/Logo corpus. The corpus profile of the Tagesschau/Logo$_{\frac{1}{5}}$ corpus was aligned with the GEO/GEOlino$_S$ corpus, (see Section 5.3.1.1).[7] When repeating the classification experiment using this data, the GEO/GEOlino$_4$ still generalized exceptionally well to Tagesschau/Logo ($acc. = 99.2\%$) whereas Tagesschau/Logo$_{\frac{1}{5}}$ only slightly improved in performance

[7] This follow-up experiment was conducted after submitting the camera-ready version of Weiss and Meurers (2018). It is therefore not reported in the original paper. However, the experiment is part of the associated poster that was presented at COLING 2018. It can be found in the online supplementary material to this thesis (`https://osf.io/5vb2x/?view_only=6d1bb8ccfe3f458c946ff4fd6ef5206b`)

on GEO/GEOlino$_S$ ($acc. = 99.2\%$). This rules out corpus size as a potential explanation for the performance difference. In the light of this finding, we propose to link the observed performance differences to the difference in the average feature merits that we discussed earlier. The average merit of features in GEO/GEOlino$_S$ was considerably lower than the average merit of features in Tagesschau/Logo. This means that the linguistic signal differentiating GEO and GEOlino texts was much weaker than the linguistic signal differentiating Tagesschau from Logo texts. In this case, it would make sense that a classifier that successfully learned to identify the weaker signal from GEO/GEOlino$_S$ will more easily react to the stronger signal provided by Tagesschau/Logo – assuming that both corpora mostly differ in terms of their signal strength not in the types of features contributing to the signal. Following this hypothesis, a classifier trained using the much stronger signal on Tagesschau/Logo instead, would likely fail to successfully notice the weaker signal on GEO/GEOlino$_S$.

5.3.3 Multi-level German L2 readability assessment

After establishing that broad linguistic complexity modeling is highly beneficial for ARA (both in terms of classification performance and explainability of results), we extended the approach to multi-level ARA for L2 readers. In Weiss *et al.* (2021), we developed an ARA model identifying texts for beginning, intermediate, and advanced L2 readers of German and English. The research aim of this work was three-fold: First, to obtain a successful multi-level readability classifier for L2 readers of German and for L2 readers of English. Much less research has been conducted on ARA for L2 readers than for L1 readers, especially for German (see Section 4.3). Furthermore, for German most research has been restricted to binary classifications (see Section 4.3), which is somewhat limited in its applicability in education contexts where a more fine-grained distinction is often necessary. Second, to investigate the cross-lingual generalizability of linguistically informed ARA models to better understand the limits and potentials of feature-based ARA in contrast to neural approaches (see discussion on multi-lingual ARA in Section 2.3.3). Third, to understand how texts at different reading levels differ from each other in terms of their linguistic complexity and whether or not there are similarities in how reading differences are expressed linguistically across languages.

In order to realize these research goals, which aimed to promote both ARA for German and for multilingual ARA, we compiled a new multi-level multi-lingual L2 readability corpus consisting of comparable articles written for L2 readers of English and German, the Spotlight corpus (see Section 5.3.1.2). The German sub section of the corpus, Spotlight-DE is the first

multi-level document-level readability corpus for German. From this data, we extracted 312 features covering all feature domains discussed in Section 2.1.2 that were applicable for both English and German. We used the multilingual CTAP system introduced in Section 3.3. 301 features were sufficiently variable across data sets and used to train two ORF classifiers with 10-CV, one for English and one for German.[8] Both classifiers were additionally evaluated in a type of zero-shot learning experiment on the sub corpus for the respective other language. We then identified and compared the most informative features for the distinction of reading levels on both Spotlight sub corpora to gain a better understanding of how reading level differences were realized linguistically for both languages. All analyses were conducted using the statistical programming language R (R Core Team, 2022) and RStudio (RStudio Team, 2022). We used the packages `caret` (Kuhn, 2022), `kernlab` (Karatzoglou *et al.*, 2022), `e1071` (Meyer *et al.*, 2022), `ranger` (Wright and Ziegler, 2017), `tidyverse` (Wickham *et al.*, 2019), `randomForest` (Liaw and Wiener, 2002), and `ordinalForest` (Hornung, 2021).

Classifier and cross-language evaluation

The German classifier achieved an accuracy of 88.0% against a majority baseline of 52.8%. Further analyses of reading level-wise F1 score, precision, and recall revealed the performance to be balanced across readability levels. Although a within-language cross-corpus evaluation as in Weiss and Meurers (2018) was not feasible due to the lack of available reference data for German, the high accuracy suggests that the model is suitable for use in practice. The German classifier also generalized to some degree to the English data ($acc. = 53.4\%$ against a majority baseline of 46.5%) but the drop in performance accuracy was considerable. Similarly, the English classifier achieved an accuracy of 74.5% on the English data which was again balanced across reading levels. This can also be considered highly successful. However, its performance dropped to 55.5% on the German data. This was still significantly higher than the majority baseline ($p = .02$ as per a one-sided t-test) but not by much. Given the linguistic differences between English and German, it is remarkable that the classifiers generalized even to this limited degree, indicating that there is some universal principle underlying the adaptation of texts to different reading skills even across languages. This link was further explored in the following study on feature informativeness.

[8] Weiss *et al.* (2021) also established the competitiveness of the complexity-feature based approach with other ARA approaches for English by achieving state-of-the-art performance on the OneStopEnglish corpus (Vajjala and Lučić, 2018) which has been used as reference data set for English ARA. However, since the focus of this thesis was on German, this was not discussed in more detail here.

Feature informativeness

We used the correlation-based feature subset selection for machine learning approach by Hall (1999) to identify the most informative features for the distinction of reading levels on Spotlight-DE and Spotlight-EN. This method allowed us to also consider the correlation between features in the ranking. The results showed that up to 32% of features were shared between the set of informative features for both languages. Especially text length and language use features were highly informative on both data sets: Text length systematically increased with higher reading levels for both languages. Language use became more variable and sophisticated. However, we also observed clear differences, which is in line with the low generalizability of models observed in the previous study. Lexical complexity played a role for both languages but not with respect to the same features. Syntactic complexity was much more informative for German than for English, whereas morphological complexity and discourse seemed to play a more central role for English. As in Weiss and Meurers (2018), human language processing measures were not relevant for the distinction of reading levels, which was unexpected given that they are motivated by psycho-linguistic theories of human sentence processing. Overall, the findings in any case confirm that the Spotlight publisher adapted texts designed for readers at different proficiency levels across a broad range of linguistic features similar to the publishers for German media language studied in Weiss and Meurers (2018). This stands in contrast to the lack of systematic adaptation observed for the publishers for content-matter school textbooks in Berendes *et al.* (2018).

5.3.4 Assessing sentence readability for German language learners

The two preceding articles in this section have established that broad linguistic complexity modeling is highly beneficial for ARA for German L1 readers and German L2 readers when analyzing longer reading texts. In Weiss and Meurers (2022), we moved from the level of full texts to the assessment of individual sentences. With this, we pursued two main research goals: First, to build a successful sentence readability classifier for German L2 readers. As Weiss *et al.* (2021), this worked towards addressing the need for more ARA models for L2 learners. It also addressed the lack of ARA approaches for the sentence level (see, e.g., Collins-Thompson, 2014, and Section 4.3). Sentence level ARA has many potential application domains inside and outside education ranging from the analysis of short social media data (such as tweets) to the evaluation of questionnaire items or the analysis of exercise de-

scriptions or captions in text books. Second, to gain a better understanding of when statistical approaches to ARA are necessary and when easier to compute traditional readability formulas suffice in practice. Statistical approaches have repeatedly been demonstrated to outperform traditional formulas (Benjamin, 2012; Collins-Thompson, 2014; Vajjala, 2022). Yet, the question remains relevant given that readability formulas have remained the *de facto* standard in research using ARA in practice due to their accessibility and ease of computation. We used the TextComplexityDE corpus (Naderi *et al.*, 2019b) to train a sentence readability regression model with 10-CV for German L2 learners that predicted readability on a 7-point Likert scale (see Section 5.3.1.2). For this, we extracted all 543 complexity features for German using the CTAP system (see Section 3.3), 373 of which were sufficiently variable on the data to inform the regression model.[9] We evaluated the model's performance for two tasks (prediction and ranking) and compared it against a regression model using only surface length features and four readability formulas: *Wiener Sachtextformel* (Bamberger and Vanecek, 1984), *Amstad readability index* (Amstad, 1978), *LIX* index (Björnsson, 1983), and the Miyazaki EFL readability index (Greenfield, 1999, 2004).

The paper also pursued a third (secondary) research goal: to use the sentence wise readability model to analyze how document level readability is being constructed. This has direct implications for automatic and manual text adaptation, yet, little previous work has addressed this issue. Vajjala and Meurers (2014) reported initial evidence that simplified and regular texts do not systematically differ in terms of their difficulty on the sentence level. However, their evidence was based on the comparison of Wikipedia and Simple Wikipedia data and seeing that the validity of Simple Wikipedia has been called into question (e.g., Štajner *et al.*, 2012; Xu *et al.*, 2015; Yaneva *et al.*, 2016), more research on other data sets was needed to verify their findings. To do so, we applied the linguistically informed sentence readability model on the Spotlight-DE corpus (see Section 5.3.1.2) to compare the differences between easy, medium, and advanced articles. All analyses were conducted using the statistical programming language R (R Core Team, 2022) and RStudio (RStudio Team, 2022). We used the packages `tidyverse` (Wickham *et al.*, 2019), `caret` (Kuhn, 2022), `data.table` (Dowle and Srinivasan, 2021), `MLmetrics` (Yan, 2016), `lattice` (Sarkar, 2008), `leaps` (Lumley, 2020), `monomvn` (Gramacy *et al.*, 2022), `krls` (Ferwerda *et al.*, 2017), `readxl` (Wickham and Bryan, 2022), `ggsignif` (Constantin and Patil, 2021), and `rstatix` (Kassambara, 2021).

[9] Note that the reduction of features by 31.3% is to be expected for short data. This could already be observed in our analysis of short answers of German L2 learners (Weiss and Meurers, 2021), which I presented in Section 5.2.4.

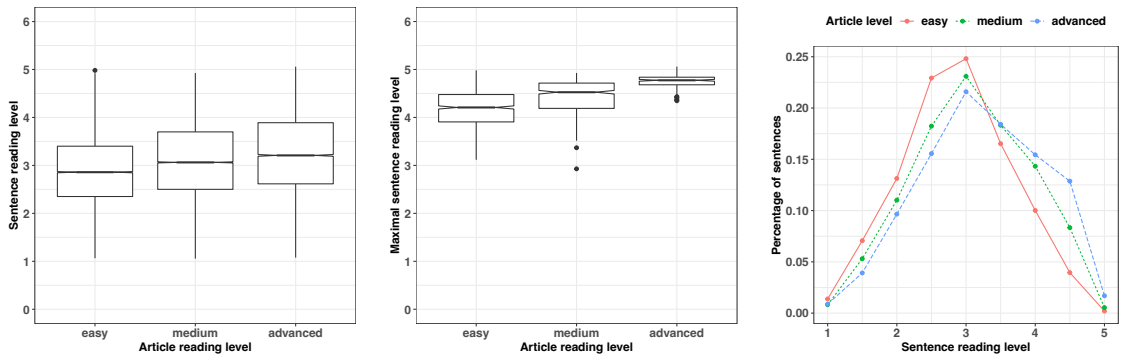
5.3.4.1 Evaluation on TextComplexityDE

For the predictive regression task, the complexity feature-based regression model trained with 10-CV showed a prediction error of a little more than half a point on the 7-point scale ($RMSE = .685$). This was considerably lower than the model informed only by surface length features ($RMSE = .739$) and the previous SOTA on the data set ($RMSE = .847$) by Naderi *et al.* (2019a). Since neither of the readability formulas predicted the same scale as human readability estimates in the TextComplexityDE corpus,we evaluated their performance using the Spearman rank correlation (r_s). Even though the formulas showed a relatively high absolute correlation ranging from .52 to .681, the complexity based model correlated considerably higher with the human estimates ($r_s = .806$). The same held for the length-based model ($r_s = .785$). For the simpler task of ranking simplified sentence pairs from the TextComplexityDE sentence pair sub corpus, the linguistically broadly informed regression model achieved a ranking accuracy of 96.0%. This is comparable to the ranking accuracy of the Amstad readability index ($acc. = 95.6\%$) and the Miyazaki EFL readability index ($acc. = 96.8\%$). It is worth noting that also the other approaches achieved high accuracy values around 93.0% and that all approaches successfully distinguished between weak and strong simplifications.

Taken together, the results demonstrated that broad linguistic complexity modeling achieves SOTA results for the task of predictive readability assessment. It outperforms traditional readability formulas and the solely surface feature based model on this task. However, for the simpler task of readability ranking for meaning equivalent pairs of regular and simplified sentences, the findings are different. Also some traditional readability formulas performed at the same level as the computationally more costly linguistically informed model for sentence ranking. Taken together, this highlights that traditional readability formulas are indeed suited for precise readability estimates. However, they can be sufficient for the distinction of artificially simplified sentences from their non-simplified counterparts.

5.3.4.2 Text profiles on Spotlight-DE

After establishing the complexity feature-based regression model as the most successful ARA model for predicting sentence readability, we applied it to predict the readability of all sentences in the Spotlight-DE corpus (see Section 5.3.1.2) which we already analyzed previously in Section 5.3.3 (Weiss *et al.*, 2021). Figure 5.1 visualizes the findings from three perspectives. Figure 5.1a shows the notched box plots of the predicted sentence-wise readability



(a) Overall difficulty (b) Maximum difficulty (c) Difficulty distribution

Figure 5.1: *Sentence difficulty profiles on Spotlight-DE across article levels*

of all sentences in Spotlight-DE for the three article reading levels easy, medium, and advanced. While there was a systematic increase in the predicted sentence difficulty score with increasing article level, the difference was not particularly pronounced. This changed in Figure 5.1b, which only considered the maximum predicted sentence readability score for each article. This suggested that maximum sentence readability was more indicative for the distinction of article readability than the sentence-wise readability score for all sentences in the article. Figure 5.1c further corroborated this finding by showing the percentage of sentences falling within a certain readability bin split by article level. The figure shows that for easy, medium, and advanced articles, more than 20% of sentences were at a medium difficulty level. However, medium and advanced articles contained much more difficult sentences than easy texts. These findings indicate that it is the maximum sentence difficulty that determines text readability rather than the average sentence readability, thus corroborating earlier findings by Vajjala and Meurers (2014) on Wikipedia and SimpleWikipedia. This has direct implications for publishers and others seeking to adapt texts to the competence of readers either by enhancing or decreasing the difficulty of texts. Rather than focusing on changing the readability level of all sentences, it might suffice to adjust the readability level of few sentences. However, more research is needed to fully comprehend the complex interplay between sentence-wise and document readability level.

Chapter 6

Conclusion

6.1 Summary of findings and limitations

With this dissertation, I have presented a new integrative approach to complexity modeling of German that can be flexibly extended to other languages and covers a linguistically broad range of measures. It integrates perspectives from different research disciplines such as SLA, computational linguistics, and psychology. This thesis, thus, takes an inherently interdisciplinary perspective. I combined insights, methods, and research questions from linguistics, machine learning, and psychology to support an integrative view on the automatic assessment of linguistic complexity for education contexts. The studies that are part of this thesis were published in computational linguistic outlets as well as in outlets for SLA research to maximize the interdisciplinary impact of this work. Furthermore, the resulting resources and methods have already successfully been used to address research questions from education science (Riemenschneider *et al.*, 2021; Weiss *et al.*, 2019), history didactics (Bertram *et al.*, 2021; Kühberger *et al.*, 2019) and German linguistics (Weiss *et al.*, 2022) in work going beyond the scope of this thesis. This demonstrates that the methodology and resources that I developed in this thesis have substantial potential to foster interdisciplinary research on language and content-matter learning and teaching.

I have further addressed important challenges in research on APA and ARA and broadened the SOTA for German. In the following, I will summarize the core findings and contributions as well as the limitations of the work presented in this thesis divided into complexity research (Section 6.1.1), APA (Section 6.1.2), and ARA (Section 6.1.3).

6.1.1 Linguistic complexity research

We have proposed a linguistically broad approach to complexity modeling for German that covers the domains of syntax, lexicon, morphology, language use, human processing, and discourse. This approach is not only linguistically broad in the sense that it expands the linguistic domains studied. It is also integrative in that it combines features, methods, and concepts from different research areas. In particular, this includes SLA complexity research as well as computational linguistic, psycho-linguistic, and psychological work on readability, discourse comprehension, language processing, and proficiency assessment. This allowed us to study the developmental variation of complexity in learner writing from a linguistically rich perspective. It also helps to overcome the reductionist approach to SLA complexity research that focuses on assessing few measures of syntactic and lexical complexity. This is an important contribution to SLA complexity research which has articulated the need to view complexity as a multi-dimensional construct and to consider the developmental variation of linguistic domains beyond syntax and lexicon (Housen *et al.*, 2019; Kuiken *et al.*, 2019; Norris and Ortega, 2009). In our effort to do so, we followed an explorative approach that views different complexity domains simultaneously.

A clear limitation of this approach is that within our studies we cannot discuss the developmental trajectory of each complexity measure in detail. When considering hundreds of measures, a detailed discussion of individual measures is simply not feasible. This is not to say that our studies did not consider the linguistic characteristics in our data. We zoomed in on a typically data-driven selection of features to better understand their developmental patterns. We also aggregated measures by their linguistic domain to gain general insights into differences across linguistic domains, as making these observable is one of the unique strengths of our linguistically broad approach to complexity. However, most measures that we assessed could not be discussed in detail due to space limitations and we could not yet investigate the lack of developmental variation in individual measures which we would have expected to vary. This is a necessary trade-off for the empirically uniquely broad view of learners' language development. It highlights that the approach advocated here is not intended as a substitute but rather as a supplement to the more common theory-driven complexity studies tracking the developmental trajectory of individual measures. It allows to gain insights that can then in turn inform experimental studies. Our linguistically broad approach was enabled by the automatic calculation of measures and by the use of feature-based machine learning algorithms. Our empirical studies demonstrated that such a linguistically broad approach is very successful for

a characterization of developmental variation that generalizes to new data and beyond task effect (see Section 6.1.2). This is a valuable methodological and conceptual transfer between SLA and computational linguistic research.

A second limitation of our approach is that the automatic calculation of complexity measures comes with a certain loss of control. Unlike with manual analyses, linguists cannot easily decide during annotation how to annotate certain non-standard language phenomena. This can be an issue when working with learner language. There is a clear trade-off to be considered here. While NLP avoids inconsistencies and increases the reproducibility of analyses, it also risks the systematically erroneous interpretation of learner data. To gain better insight into how prevalent this problem was in our studies, we evaluated the performance of our system on a gold standard annotated L2 data set. Our analyses showed that linguistic units could be determined with high accuracy and that erroneous analyses had little impact on most measures. While these are encouraging findings for the presented studies, the system should be evaluated further in future work. Our findings are based on a small set of short responses from beginning learners of German. Other learner corpora with leveled language productions and gold standard annotations of a wide range of linguistic constructs were not available for German at the time of writing. When such data becomes available, the evaluation should be extended to longer productions and and a wider range of proficiency levels to validate the results. It would also be worthwhile to study the robustness of our approach on different types of non-standard language use.

Despite these limitations, our approach proved to be a valuable contribution to complexity research on German. To make this broad range of measures accessible to other researchers, I integrated the analysis system into CTAP, a web platform that was originally designed for the analysis of English complexity. The web-based graphical user interface grants access to a uniquely rich set of complexity measures to users without the technical background or resources to implement these themselves. At the same time, this supports the comparability and reproducibility of complexity research as researchers can use the same analytical resources across studies. While integrating the German analysis system into CTAP, I extended and aligned the existing collections of measures for English and German. In this context, I adapted the system architecture to facilitate a shared pipeline for different languages to avoid redundancies and promote the addition of new languages. The success of this design has been demonstrated by the ongoing addition of new languages into the system running at `www.ctapweb.com`, including French, Spanish, Dutch, and Portuguese. Derivatives of CTAP

have also been created for Chinese (Cui *et al.*, 2022) and Italian (Okinina *et al.*, 2020).

6.1.2 Automatic proficiency assessment

Our systematic survey on ALPS research for German characterized the research landscape as scarce in the sense that only a total of 23 papers fully matched our inclusion criteria. This highlights the need for more research on APA (or ALPS in general) for German. Turning to the included papers, we found that research on ALPS has been conducted across research disciplines related to education and that there was a successful methodological transfer of machine learning techniques from computer science and computational linguistics to other disciplines such as SLA research and psychology. However, we also observed that cross-corpus studies play virtually no role in the evaluation of German ALPS models. Furthermore, most research focused on adults and the evaluation of L1 writing performance on longer texts. Against this background, all three papers on APA that are part of this thesis make an important contribution to addressing the lack of work on assessing long and short L2 writing and early L1 academic language development from elementary to early secondary school. We further found that measures of discourse complexity play only a minor role in German approaches to ALPS which is in contrast to the relevance of discourse measures in English research on AWE and ATS (see Crossley, 2020). I reasoned that this might be caused by the lack of automatic tools that support the analysis of discourse complexity, which highlights the importance of making such measures available via CTAP.

In terms of linguistic developmental trajectories, our research showed that linguistic domains developed independently from each other and differed in their importance for the characterization of different proficiency levels. We observed that L2 writing was characterized across the full CEFR range by changes in the lexical and clausal domain whereas other linguistic domains—such as discourse and morphology—were important for the characterization of specific proficiency levels. Similarly, our analysis of early L1 academic language acquisition revealed differences in the development of linguistic domains. Accuracy proved to be particularly relevant for the distinction of grade levels in elementary school writing whereas writing in early secondary school developed primarily in terms of phrasal, lexical, and discourse complexity. In short, we found that our integrative approach to linguistically broad complexity modeling is beneficial for L2 and L1 proficiency assessment on longer texts. Our analysis of L1 data further demonstrated the cross-prompt generalizability of our models by testing them on unseen task prompts. This not only demonstrated the quality of the result-

ing machine learning model—an important concern in machine-learning-based approaches to ALPS. It also addressed the concern of unaccounted task variation biasing the developmental variation that we were interested in. In contrast, for short L2 writing, we found a limited generalizability of the models, likely due to the limited linguistic evidence that can be retrieved from short answers.

Our study of L2 short answers also demonstrated the robustness of our approach on non-standard data. Seeing that our analyses are fully automated, evaluating the performance of our feature extraction algorithm on non-standard language data is central to understand the reliability and interpretability of our linguistic insights. We saw that our NLP pipeline performs well on beginning learners' short answers and that errors in the analysis have a limited impact on the calculation of complexity features. It should be noted though that we analyzed relatively simple sentences. Ideally, the analysis should be extended in future work to syntactically more elaborate language to see if this impacts the calculation of syntactic complexity measures more heavily. Unfortunately, this was not feasible for the present thesis because of the lack of learner corpus data that contains linguistic annotations, as well as proficiency annotations and is sufficiently large to train an APA model.

Generally, the reliance of the presented studies on available corpora that are suited to train APA models in terms of their annotations and size was one of the main limitations of the presented work. The corpora we used differed in terms of their proficiency annotations (expert annotations, grade levels, and course levels) and elicitation contexts. This makes it difficult to directly compare our findings regarding the developmental variation of complexity in L2 and L1 writing. That being said, we made the following seemingly parallel observations in the developmental trajectories of L1 and L2 writing: We found that lexical complexity and language use as well as features of discourse cohesion were consistently among the most important indicators of developmental variation for both L1 and L2 writing. Also, syntactic complexity played an important role, even though L2 proficiency developed more in terms of clausal complexity whereas phrasal complexity was more important for L1 proficiency. Morphological complexity increased for both L1 and L2 writing, albeit with regard to different features.

Finally, while approximating proficiency through course levels is common practice in research on APA and SLA complexity research, it is a coarse approximation of language proficiency which coincides with related but conceptually distinct aspects such as age and duration of exposure to language or instruction. However, large annotated learner corpora for German

are rare and at the time of writing, these were the best options for our research purposes. Our research goal was to develop a linguistically broad and interdisciplinary approach to automatic complexity assessment and to present its advantages for the analysis of L1 and L2 data in educational contexts. This would not have been feasible without relying on existing corpus resources. We regard the associated limitations as acceptable trade-offs considering the lack of comparable work for German. This allowed us to train several high-performing models for APA for German. This includes models for L1 and L2 speakers on long text productions and—for L2 speakers—on the level of short answers (less than 10 words).

6.1.3 Automatic readability assessment

Reviewing the research landscape of ARA research for German highlighted that ARA is being used across research disciplines, i.e., also outside of education contexts, for example in work focusing on web accessibility. In contrast to our observations in the ALPS survey, though, we saw a clear methodological divide across research disciplines in the sense that machine-learning-based approaches play virtually no role outside of computational linguistic work on ARA. Readability formulas remain the de facto standard across research disciplines. This is an important insight with clear implications for future research on ARA for German. Against this background, we critically compared the performance of readabilty formulas and linguistically rich machine-learning-based approaches to ARA across a variety of ARA tasks. Our findings demonstrated that while readability formulas are suited to produce coarse-grained relative readability estimates in sentence pair ranking, SOTA methods achieve considerably better performances when it comes to the prediction of readability levels. This showcases that the use of readability formulas in practice is an issue that should be addressed with models that are more accessible. This thesis has partially contributed to this by making the models and analysis tools available online. All three ARA models set a new SOTA for readability assessment on the respective corpora.

Our survey further showed that ARA has mostly focused on adult L1 readers as a target population. Yet, we also found individual contributions focusing on a wide array of target populations including children and L2 readers. In this respect, the ARA studies presented in this thesis tie in with the less represented target populations for German ARA: L2 readers and children. Furthermore, little research has focused on the readability of sub-textual units such as sentences. This makes our work on sentence-level readability assessment a particularly relevant contribution to German ARA research, also because it allows insights into the

compositionality of document-level readability. We found evidence that maximal sentence difficulty rather than the average readability of sentences within a document determines its overall readability. This has important implications for research on text simplification. More research is needed to elaborate on the link between sentence and document readability, both to promote work on discourse comprehension and text simplification.

Finally, our systematic review showed the need for more cross-corpus validation of machine-learning-based ARA models. The lack of such studies is partially due to the limited resources for German ARA and the lack of available corpora is a central limiting factor for ARA in general and for German in particular (Collins-Thompson, 2014; Vajjala, 2022). We addressed this need by compiling three new readability corpora for German in the context of this thesis. We compiled the new GEO/GEOlino corpus and the Tagesschau/Logo corpus (Weiss and Meurers, 2018) that both represent language use in expository German media language targeting adults and children (6 to 14 years). Both corpora represent authentic language use and support cross-corpus testing of ARA models for the comprehensibility of media language. A limitation of Tagesschau/Logo and GEO/GEOlino is that they only support the binary distinction between adults and children. However, combined they allow for cross-corpus evaluation of models for the readability of German media language and they were at the time of writing the only available L1 readability corpora for German that can be used to identify materials for children. Thanks to these two corpora, we were able to confirm the cross-corpus generalizability of our models and gain linguistic insights into the use of German media language. We also compiled the Spotlight corpus (Weiss *et al.*, 2021) which is a multi-lingual corpus of leveled reading materials for non-native readers of English and German. It is the first large multi-level readability corpus for German and one of the first multi-lingual multi-leveled readability corpora. All corpora are available upon request for research purposes.[1]

Turning to the linguistic insights that we gained through our studies, we found that features of language use, nominal style, and discourse complexity were particularly informative for the distinction of media language targeting adults and children. For L2 readability, we found syntactic complexity, language use, and surface length measures to be particularly important. Language use and surface length were not only central to the distinction of readability levels for German but also for English. We were able to make this comparison because we compiled the multilingual multi-level Spotlight readability corpus for L2 readers that currently supports German and English and is being extended to Spanish, French, and Italian. Even though we

[1]Contact dm@sfs.uni-tuebingen.de or zarah-leonie.weiss@uni-tuebingen.de for access.

observed only a limited generalization of our language-specific models when applying them to new languages, our comparison provides us with important insights regarding the similarities and differences between leveled German and English reading materials for L2 learners.

Despite the overall success of our integrative approach to linguistic complexity modeling for ARA, our work has some important limitations. The annotation validity of the readability corpora that we compiled has so far not been confirmed experimentally: The corpora Tagesschau/Logo and GEO/GEOlino infer the readability level from the target group of the published media. The Spotlight corpus contains leveled articles from the Spotlight publisher. The publisher equates these articles to the CEFR levels A2, B1/B2, and C1. In contrast, the TextComplexityDE corpus (Naderi *et al.*, 2019a) that we used to study sentence-level readability was based on aggregate scores of experimentally elicited human readability judgments. For our three leveled readability corpora, it is unclear how publishers adjusted materials to their intended target groups, for example if they used specific guidelines, expert judgments, or readability formulas of their own. This is not to say that we have not considered the reliability of out reference annotations. As Vajjala (2022) pointed out, labels provided by professional publishers of educational materials have a certain credibility and their use is a standard procedure in ARA research. We further reasoned that the economic success of the outlets for children and L2 learners indicates some success in their alignment to the target population. The generalizability of our models across corpora further confirmed this intuition because the generalization to data from other publishers demonstrates that our models were not only based on idiosyncratic differences made by individual publishers. Despite this encouraging evidence in favor of the validity of the annotations, it remains important to also systematically validate the readability annotations through reader experiments. Unfortunately, this was not feasible in the context of this thesis given that we wanted to support cross-corpus and cross-lingual testing for our ARA models. Thus, experimentally confirming the annotation validity of the corpora we compiled for this thesis and procuring more would be central to promote German ARA research.

For the same reasons, we could not address individual reader characteristics in the context of this thesis. Experimentally collected readability labels that encode reader characteristics as meta-information do not currently exist for German in corpora that are suitable for training ARA models. Therefore, we could not consider such factors in the context of this dissertation, despite their importance for discourse comprehension.

6.2 Implications for research and teaching practice

In this thesis, I have focused on advancing SLA complexity research and computational linguistic research on APA and ARA by presenting a linguistically broad, integrative approach to automatic complexity modeling for German. However, the approach presented in this thesis has also important implications for education research in other disciplines and teaching practice. Even when language is not the subject matter itself, it is the central medium of human communication and as such plays a crucial role in content-matter learning and teaching. As such, the analysis of language performance and comprehension has relevance in education beyond research on language learning. Important application domains outside of language learning and teaching can for example be the analysis of how comprehensible language input is in content-matter teaching. In subject matters that elicit open responses from learners, the linguistic performance of learners can also play a role, especially when it comes to assessing the register-awareness of learners for their subject matter. Shifting our perspective away from learners, we can also use complexity analyses to analyze the performance of language and content-matter teachers.

These are not purely theoretical considerations on the potential future use cases for the approach to complexity modeling presented in this thesis. Paralleling the research that has contributed to this dissertation, we have systematically worked on the application of our approach to authentic data from education contexts and interdisciplinary exchanges with other fields of education-related research. To include these in the thesis would have went beyond the scope of this dissertation. However, a list of all articles dedicated to German language modeling in applied education settings can be found below. In the following, I briefly outline the work done in these articles that focuses on complexity modeling, both for the sake of completeness and as examples of the broader application potential of the approach we presented.

1. Bertram, C., Weiss, Z., Zachrich, L., and Ziai, R. (2021). Artificial intelligence in history education. Linguistic content and complexity analyses of student writings in the CAHisT project (computational assessment of historical thinking). *Computers and Education: Artificial Intelligence*, page 100038.

2. Dittrich, S., Weiss, Z., Schröter, H., and Meurers, D. (2019). Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), Turku Finland*, pp. 41–56.

3. Kühberger, C., Bramann, C., Weiss, Z., and Meurers, D. (2019). Task complexity in history textbooks: A multidisciplinary case study on triangulation in history education research. *History Education Research Journal*, 16.

4. Riemenschneider, A., Weiss, Z., Schröter, P., and Meurers, D. (2021). Linguistic complexity in teachers' assessment of German essays in high stakes testing. *Assessing Writing*, 50.

5. Weiss, Z., Dittrich, S., Schröter, H., and Meurers, D. (2019). A Linguistically-Informed Search Engine to Identifiy Reading Material for Functional Illiteracy Classes. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, 7th November 2018*, pp. 79–90.

6. Weiss, Z., Lange-Schubert, K., Geist, B., and Meurers, D. (2022). Sprachliche Komplexität im Unterricht. Eine computerlinguistische Analyse der gesprochenen Sprache von Lehrenden und Lernenden im naturwissenschaftlichen Unterricht in der Primar- und Sekundarstufe. *Zeitschrift für Germanistische Linguistik*, 50(1), pp. 159–201.

7. Weiss, Z., Riemenschneider, A., Schröter, P., and Meurers, D. (2019). Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 30–45.

6.2.1 Proficiency and readability assessment in history teaching

In this thesis, I have presented systematic approaches to ARA and APA in L2 and L1 writing. The models trained on L2 data focused on data from language learning contexts whereas the models trained on L1 data focused on general academic language competencies. However, writing skills and reading comprehension also play a crucial role in content-matter teaching, especially in humanities subjects where subject matter appropriate language use and open-ended answer formats play a particularly important role. Applying the presented approach to such a context is a natural extension of the presented work. We chose history as an example subject matter to investigate the transferability of our approaches to new teaching domains.

Paralleling our work on ARA, we analyzed the complexity of history tasks in Austrian textbooks in Kühberger *et al.* (2019) from an interdisciplinary, triangulative perspective. We operationalized 'task complexity' from an interdisciplinary perspective, considering three com-

ponents: i) general task complexity, ii) general linguistic complexity, and iii) domain-specific task complexity. The approach presented in this thesis was used to approximate general linguistic complexity. This was also our first attempt to estimate the comprehensibility of very short language samples. Since there continues to be no multi-level model for capturing readability for L1 readers for German—the model presented in this thesis adopts a binary division between adults and children—we used the proficiency model from Weiss and Meurers (2019a) for the study. We thus evaluated the correspondence between the linguistic complexity of the task prompt and the typical language production competence of pupils at the corresponding grade level. Despite this limitation, the approach showed promising interactions with the other components of task complexity. However, replicating the study with a more appropriate model would be a desirable goal for future research.

In Bertram *et al.* (2021), we explored the feasibility of automatically assessing the quality of pupils' answers to open-ended history tasks in an interdisciplinary collaboration between computational linguistics and history didactics. We combined computational linguistic analyses of automatic content scoring for short answers (using the approach by Ziai, 2018) with our approach to linguistic complexity modeling. This allowed us to consider task-appropriate language use in addition to the factual correctness of answers. The data we analyzed had been elicited to study pupils' history competencies through answers to prompts requiring increasingly sophisticated historical reasoning of learners. Our exploration yielded promising first results and provided evidence that historical thinking requirements impact pupils' language performance in terms of its linguistic complexity. We consider this to be an important first step towards supporting history teaching with computational linguistic tools. However, parallel to current practice in ATS (Attali, 2007; Powers *et al.*, 2002; Zhang, 2013), we recommend that such a system be used as a complement to, rather than a replacement for, human raters, especially in high-stakes contexts.

6.2.2 Assessing teachers' grading objectivity and classroom language

This thesis took a learner-centered perspective on complexity analysis in education contexts. We analyzed the language production of learners and reading materials for learners. However, another relevant focus for our analysis are teachers and their reception of learner language as well as teachers' own language productions in authentic teaching and learning contexts. We applied our approach to study these aspects in more detail in three studies.

In Weiss *et al.* (2019) and Riemenschneider *et al.* (2021), we adapted our approach to lin-

guistic complexity modeling to evaluate teachers' grading objectivity in high-stakes testing, namely in the evaluation of German *Abitur* writing for the subject matter German.[2] In Weiss *et al.* (2019), the goal of our complexity analysis was the identification of 16 texts that are comparable in their overall performance but maximally differ in their overall language performance. These texts were used for a subsequent study on teachers' grading objectivity and the potential influence of linguistic complexity and accuracy on their judgments. To identify suitable materials from a pool of 344 essays, we used our analysis system to represent each essay as a complexity vector. Rather than using all complexity measures, we selected relevant measures through a combination of theory- and data-driven feature selection. I then calculated the correlation between complexity measures and the grades that they had received in the *Abitur*. I used this information to build artificial vectors of ideal language use. We did so independently for each of the four different task prompts represented in the data to account for task-appropriate language use.[3] We obtained essays that were maximally different in their language performance by selecting essays of comparable length whose complexity vectors were maximally close (or maximally distant) from the ideal complexity vector calculated for their task prompt. We only considered essays for this step that had received a medium overall grade to ensure the comparability of essays in terms of their content quality and to avoid flooring and ceiling effects in the subsequent study. The selected essays were error corrected to be able to distinguish between the influence of complexity and accuracy and used as corrected and non-corrected versions for the study.

After this preparation, our collaborators at the Institute for Educational Quality Improvement (IQB, Berlin) conducted a rating experiment with 33 experienced teachers to re-rate the selected essays. The results revealed that teachers successfully recognized linguistic differences between essays when asked to grade them on their linguistic performance. We also found no undue influence of complexity on teachers' content grades. However, accuracy showed to impact teachers' judgments of content quality. We elaborated on these analyses in Riemenschneider *et al.* (2021) to gain more insights into the interactions between the linguistic complexity of students' writing, teachers assessment of the complexity of students' writing, and teachers grading of students' writing.

Turning to teachers' language production in education contexts, we investigated teachers'

[2] In Germany, the *Abitur* is the final examination of pupils across different subject matters at the end of the academic track of secondary school (the German *Gymnasium*). It is a mandatory entry requirement for German university.

[3] A post-hoc analysis confirmed substantial differences regarding the ideal language use across task prompts.

language use in authentic classroom interactions in Weiss *et al.* (2022). This study also tests the possibility of examining transcripts of spoken language using our analysis system presented in this dissertation. We analyzed transcripts of teachers' classroom language in late elementary school (4th grade) and early secondary school (6th grade in *Gymnasium*—the academic secondary school track—and *Hauptschule*—the vocational secondary school track). The lessons focused on the condensation and vaporization of water, thus keeping the topic constant across teachers and grade levels. Our analysis showed complexity differences across linguistic domains in the language of teachers between elementary school and secondary school, but with clear differences between *Gymnasium* and *Hauptschule*. Teachers' language in *Gymnasium* is systematically more complex compared to elementary school, whereas teachers' language in *Hauptschule* is characterized by a lower complexity compared to elementary school. Whether this difference is appropriately adaptive to pupils' language proficiency cannot be determined without further individual annotations. We identified the lack of speech target annotations—who is being addressed—to be a central limitation for the analysis of adaptivity in spoken interactions between multiple interlocutors. Still the results provide evidence that academic language input in classroom interactions does not necessarily systematically increase across grade levels and school types in German content-matter teaching.

6.3 Outlook and future research directions

Concluding this thesis, I would like to briefly outline future research directions that emerge from the presented work. Beyond the natural extension of the presented approach through the addition of more complexity measures (e.g., covering the phonological domain), languages (e.g., Arabic), and training more models for different target groups (e.g., low literate readers), I identified four core research desiderata for future work: i) linking proficiency and readability assessment, ii) developing a procedural approach to complexity analyses, iii) experimentally validating the corpus annotations, feature extraction, and model predictions, and iv) making models more accessible for users without a technical background.

Let us first turn to linking proficiency and readability assessment. This thesis has focused on outlining the use of broad linguistic complexity modeling for APA and ARA in education contexts. This opens up the possibility of investigating links between the production and reception of language. For example, we observed an increase in nominal style and noun complexity in the development of early L1 academic language writing which is paralleled

by the differences between media language targeting adults and children. Furthermore, we saw that language use and discourse cohesion were relevant for both APA and ARA for L1 learners. In our studies on longer L2 writing and texts for L2 readers, we showed that the lexical and syntactic domain as well as language use measures were particularly relevant for both proficiency assessment and assessing the readability of texts. These parallels that we can observe due to the shared approach to automatic complexity modeling are a promising first step to empirically quantifying learners' ZPD (Vygotsky, 1978) or $i+1$, as the distance between learners' proficiency and the ideal input is termed in Krashen's (1985) influential 'Input Hypothesis'. For English, similar approaches have already yielded promising results for individually competence-adaptive text retrieval (Chen and Meurers, 2019) and to inform tutoring systems (Watson and Kochmar, 2021). However, due to the lack of resources and models, such work has not been feasible for German until now.

Second, throughout this thesis, we have taken a resultative perspective on proficiency and readability in the sense that we assigned a single score to a full text, focusing on reading and writing as a product. Yet, reading and writing (as well as listening and speaking) are incremental processes. This becomes particularly apparent when analyzing interactive speech or chat data which has a more obvious temporal dimension as discourse evolves and changes longitudinally. However, it is also relevant for the assessment of writing proficiency and text readability. Identifying passages of texts that are challenging within their context rather than assigning a single overall label to texts would be a desirable extension of current work on ARA. It could also facilitate the procedural assessment of writing quality can promote formative feedback. The first steps in this direction were taken within the context of this dissertation by extending our approach to short language samples. These models can be used to identify the readability or quality of a text at any given position in the text. First work in this direction has been proposed by Marcus *et al.* (2016); Ströbel *et al.* (2020) who used a moving window technique to estimate writing quality. We believe this to be an important line of research that would be beneficial to provide formative feedback during writing as well as support the targeted adaptation of reading materials for different readers. It would also promote the longitudinal analysis of the development of spoken and written discourse and alignment processes between speakers, thus naturally extending the work presented in this thesis.

Third, more work on validating the validity of annotation labels and model predictions is needed. In this thesis, I focused on assessing the quality of our predictions using within- and cross-prompt testing, including cross-corpus, cross-task, cross-language, and cross-university

testing. The results demonstrated the generalizability of most of our models. Our findings partially indicated a more limited generalizability for models trained on very short learner productions and for cross-language predictions of readability. However, ARA and APA research rarely independently validates the correspondence between model predictions and humans in practice. More work is needed that experimentally confirms that the predictions of readability models in fact provide readable materials for readers (for a similar call, see Vajjala, 2022). Experimental studies are also needed to study the interaction between predictions by ARA models and individual reader properties as well as different reading goals. This also holds for APA: even though ATS models for English are more commonly validated, we found this to be an ongoing research gap for work on German.

Finally, predictive models for APA and ARA should become more accessible for users without a technical background to broaden the impact of this work on teaching and learning practice. We observed for ARA that readability formulas continue to dominate in practice due to their accessibility and ease of use. This relativizes the significance of SOTA research on ARA as it has little to no impact on real-life usage. While this issue was less pronounced in our survey of APA, it is true that to date there is no tool for German that utilizes APA models and the linguistic insights gained from broad linguistic modeling to provide formative feedback to learners. At the time of writing, such systems only existed for English, not for German. With this thesis, we have laid the foundation for such systems, making the analysis pipeline available through CTAP and the trained models in the online supplementary material to this thesis. To further support better access to the models trained in this thesis through user interfaces, CTAP should be made accessible through an application programming interface (API). This would make it possible to integrate the trained models into learning platforms or web tools that could extract the features required for the predictions through the API. First work in this direction has been done for German, for example in the context of the search engines KANSAS (Weiss *et al.*, 2018) and FLAIR (Chinkina *et al.*, 2016), which, however, rely on readability formulas.

The contributions made in this dissertation serve as a basis for research in these directions while making important immediate contributions to SLA complexity research, and computational linguistic approaches to ARA and APA.

Chapter 7

References

7.1 Automatic language proficiency scoring papers

Arnold, T. and Weihe, K. (2016). Network motifs may improve quality assessment of text documents. In *Proceedings of the 2016 Workshop on Graph-based Methods for Natural Language Processing*, pages 20–28, San Diego, California, USA. Association for Computational Linguistics.

Bertram, C., Weiss, Z., Zachrich, L., and Ziai, R. (2021). Artificial intelligence in history education. Linguistic content and complexity analyses of student writings in the CAHisT project (computational assessment of historical thinking). *Computers and Education: Artificial Intelligence*, page 100038.

Daller, H., Hout, R. V., and Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, **24**, 197–222.

Frey, J.-C. (2020a). Age-related language in South Tyrolean social media. In *Using data mining to repurpose German language corpora. An evaluation of data-driven analysis methods for corpus linguistics*, pages 198–250. Università di Bologna.

Frey, J.-C. (2020b). Exploring holistic text quality ratings. In *Using data mining to repurpose German language corpora. An evaluation of data-driven analysis methods for corpus linguistics*, pages 86–197. Università di Bologna.

Hancke, J. (2013). *Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language*. Master's thesis, Eberhard Karls Universität Tübingen, Tübingen, Germany.

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., and Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, **3**, 897–915.

Rama, T. and Vajjala, S. (2021). Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling? http://arxiv.org/abs/2102.12971.

Riemenschneider, A., Weiss, Z., Schröter, P., and Meurers, D. (2021). Linguistic complexity in teachers' assessment of german essays in high stakes testing. *Assessing Writing*, **50**.

Stiegelmayr, A. and Mieskes, M. (2018). Using argumentative structure to grade persuasive essays. In G. Rehm and T. Declerck, editors, *International Conference of the German Society for Computational Linguistics and Language Technology*, volume 10713 LNAI, pages 301–308. Springer Verlag.

Ströbel, M., Kerz, E., and Wiechmann, D. (2020). The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning*, **70**, 732–767.

Strobl, C. (2014). Affordances of web 2.0 technologies for collaborative advanced writing in a foreign language. *CALICO Journal*, **31**, 1–18.

Szügyi, E., Etler, S., Beaton, A., and Stede, M. (2019). Automated assessment of language proficiency on German data. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 30–39.

Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2018)*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Vanhove, J., Bonvin, A., Lambelet, A., and Berthele, R. (2019). Redicting perceptions of the lexical richness of short French, German, and Protuguese texts using text-based indices. *Journal of Writing Research*, **10**, 499–525.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *Modern Language Journal*, **96**, 576–598.

Wahlen, A., Kuhn, C., Zlatkin-Troitschanskaia, O., Gold, C., Zesch, T., and Horbach, A. (2020). Automated scoring of teachers' pedagogical content knowledge – a comparison between human and machine scoring. *Frontiers in Education*, **5**.

Weiss, Z. (2017a). Modeling L2 proficiency in Falko Georgetown. In *Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects*, pages 158–178. Eberhard Karls Universität Tübingen, Tübingen, Germany.

Weiss, Z. (2017b). Modeling L2 proficiency in Merlin. In *Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects*, pages 126–157. Eberhard Karls Universität Tübingen, Tübingen, Germany.

Weiss, Z. and Meurers, D. (2019a). Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019)*, pages 380–393.

Weiss, Z. and Meurers, D. (2019b). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In A. Abel, A. Glaznieks, V. Lyding, and L. Nicolas, editors, *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference. Corpora and Language in Use*, pages 419–435, Louvain-la-Neuve. Presses universitaires de Louvain.

Weiss, Z. and Meurers, D. (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task. *International Journal of Learner Corpus Research*, **7**, 83–130.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2015)*, pages 224–232, Denver, Colorado, USA. Association for Computational Linguistics.

7.2 Automatic readability assessment papers

Andreessen, L. M., Gerjets, P., Meurers, D., and Zander, T. O. (2021). Toward neuroadaptive support technologies for improving digital reading: a passive BCI-based assessment of mental workload imposed by text difficulty and presentation speed during reading. *User Modeling and User-Adapted Interaction*, **31**, 75–104.

Aumiller, D. and Gertz, M. (2022). Klexikon: A German dataset for joint summarization and simplification. *arXiv,* `http://arxiv.org/abs/2201.07198`.

Battisti, A. (2019). *Automatic Cluster Analysis of Texts in Simplified German*. Master's thesis, University of Zurich, Zurich, Switzerland.

Battisti, A. ., Ebling, S. ., Volk, M., Battisti, A., and Ebling, S. (2019). An empirical analysis of linguistic, typographic, and structural features in simplified German texts. In *Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari. CLiC-it.

Beime, B. and Menges, K. (2012). Does the requirement of readability testing improve package leaflets? Evaluation of the 100 most frequently prescribed drugs in Germany marketed before 2005 and first time in 2007 or after. *Pharmaceutical Regulatory Affairs: Open Access*, **1**.

Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., and Trautwein, U. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, **110**, 518–543.

Betschart, P., Zumstein, V., Ali, O. H., Schmid, H.-P., and Abt, D. (2018). Readability assessment of patient education material published by German-speaking associations of urology. *Urologia internationalis*, **100**, 79–84.

Betschart, P., Staubli, S. E., Zumstein, V., Babst, C., Sauter, R., Schmid, H.-P., and Abt, D. (2019). Improving patient education materials: a practical algorithm from development to validation. *Current Urology*, **13**, 64–69.

Bischof, D. and Senninger, R. (2018). Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research*, **57**, 473–495.

Brettschneider, F., Haseloff, A. M., and Kercher, J. (2009). Kann man Wahlaussagen verstehen? Über die Sprache der Parteiprogramme zur Bundestagswahl 2009. *Bundestagswahl 2009*, **9**, 668–670.

Brück, T. V. D. and Hartrumpf, S. (2007a). A readability checker based on deep semantic indicators. In *Language and Technology Conference*, pages 323–244. Springer.

Brück, T. V. D. and Hartrumpf, S. (2007b). A semantically oriented readability checker for German. In *Proceedings of the 3rd Language and Technology Conference*, pages 270–274.

Brück, T. V. D. and Leveling, J. (2007). Parameter learning for a readability checking tool. *LWA*, pages 149–153.

Brück, T. V. D., Helbig, H., and Leveling, J. (2008a). The readability checker Delite. Technical report, Intelligente Informations- und Kommunikationssysteme Informatikzentrum Hagen, Hagen, Germany.

Brück, T. V. D., Hartrumpf, S., and Helbig, H. (2008b). A readability checker with supervised learning using deep indicators. *Informatica*, **32**, 429–435.

Brügelmann, H. and Brinkmann, E. (2021). Wie kann man erfassen, was Texte für echte Leseanfängerinnen leicht oder schwierig macht? Zur Begründung des "Bremer Erstlese-Index" (BRELIX). *pedocs – Open Access Erziehungswissenschaften*.

Brütting, J., Reinhardt, L., Bergmann, M., Schadendorf, D., Weber, C., Tilgen, W., Berking, C., and Meier, F. (2018). Quality, readability, and understandability of German booklets addressing melanoma patients. *Journal of Cancer Education*, **34**, 760–767.

Budd, R., Marius, T., Gatewood, P., and Jones, D. (2019). Using k-means in SVR-based text difficulty estimation. In *Workshop on Speech and Language Technology in Education*, pages 84–88. International Speech Communication Association.

Calafato, R. and Gudim, F. (2020). Literature in contemporary foreign language school textbooks in Russia: Content, approaches, and readability. *Language Teaching Research*, page 1362168820917909.

den Breejen, M. (2015). *Lesbarkeit im DaF-Unterricht. Ein Vergleich zwischen verschiedenen Methoden für das Bestimmen des Lesbarkeitsniveaus und Konsequenzen für den DaF-Unterricht in den Niederlanden*. Master's thesis, Universiteit Utrecht.

Dirga, R. N. and Wijayati, P. H. (2018). How can teachers assess reading skills of generation z learners in German language class? In *IOP Conference Series: Materials Science and Engineering*, volume 296, page 12026.

Dittrich, S., Weiss, Z., Schröter, H., and Meurers, D. (2019). Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 41–56, Turku, Finland.

Eisele, O., Tolochko, P., and Boomgaarden, H. G. (2021). How do executives communicate about crises? A framework for comparative analysis. *European Journal of Political Research*.

Esfahani, B. J., Faron, A., Roth, K. S., Grimminger, P. P., and Luers, J. C. (2016). Systematische Analyse der Lesbarkeit von Patienteninformationstexten auf Internetseiten von Kliniken für Allgemein- und Viszeralchirurgie deutscher Universitätskliniken. *Zentralblatt für Chirurgie-Zeitschrift für Allgemeine, Viszeral-, Thorax-und Gefäßchirurgie*, **141**, 639–644.

Fink, J. (2012). Die Verständlichkeit von Parteien in der politischen Online-Kommunikation. *Bürgerproteste im Spannungsfeld von Politik und Medien: Beiträge zur 7. Fachtagung des DFPK*, **2**, 105.

Frey, F. (2015). *Medienrezeption als Erfahrung: Theorie und empirische Validierung eines integrativen Rezeptionsmodus*. Springer VS.

Fricke, U. (2021). Internet tools for learning level-appropriate text selection in German as a

Foreign Language? *German as a Foreign Language*, **3**.

Friedemann, J., Schubert, H. J., and Schwappach, D. (2009). Zur Verständlichkeit der Qualitätsberichte deutscher Krankenhäuser: Systematische Auswertung und Handlungsbedarf. *Gesundheitswesen*, **71**(1), 3–9.

Friedrich, M. C. G. and Heise, E. (2019). Does the use of gender-fair language influence the comprehensibility of texts? An experiment using an authentic contract manipulating single role nouns and pronouns. *Swiss Journal of Psychology*, **78**(1-2), 51.

Galasso, S. (2014). Exploring textual cohesion characteristics for German readability classification. Bachelor's thesis, Eberhard Karls Universität Tübingen.

Gilg, E., Schmellentin, C., Dittmar, M., and Schneider, H. (2019). Selbstregulation beim Verstehen von Schulbuchtexten der Biologie auf der Sekundarstufe I. *Bulletin suisse de linguistique appliquée*, **109**, 129–151.

Golke, S. and Wittwer, J. (2017). High-performing readers underestimate their text comprehension: Artifact or psychological reality? *CogSci*.

Gritz, W., Hoppe, A., and Ewerth, R. (2021). On the impact of features and classifiers for measuring knowledge gain during web search - a case study. In *Proceedings of the CIKM 2021 Workshops*, Gold Coast, Queensland, Australia.

Grzybek, P. (2010). Text difficulty and the Arens-Altmann law. *Text and Language*, pages 57–70.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Harbach, M., Fahl, S., Muders, T., and Smith, M. (2012). Towards measuring warning readability. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 989–991.

Harbach, M., Fahl, S., Yakovleva, P., and Smith, M. (2013). Sorry, i don't get it: An analysis of warning message texts. In *International Conference on Financial Cryptography and Data Security*, pages 94–111.

Heim, N., Faron, A., Fuchs, J., Martini, M., Reich, R. H., and Löffler, K. (2017). Die Lesbarkeit von onlinebasierten Patienteninformationen in der Augenheilkunde. *Ophthalmologe*, **114**(5), 450–456.

Heim, N., Faron, A., Wilms, C. T., Reich, R. H., and Martini, M. (2019). Lesbarkeit von onlinebasierten Patienteninformationen in der MKG-Chirurgie. *Der MKG-Chirurg*, **12**(3),

154–159.

Herzberg, A. (2018). *Analyse der oberflächlichen Merkmale von Qualitätsjournalismus-Texten*. Master's thesis, Hochschule für angewandte Wissenschaften Hamburg, Hamburg, Germany.

Hewett, F. and Stede, M. (2021). Automatically evaluating the conceptual complexity of German texts. In *Proceedings of the 17th Conference on Natural Language Processing*, pages 228–234.

Hinz, M. (2020). *Verständlichkeit von Gebrauchsanleitungen: Sprachlich-rechtliche Anforderungen an Produkt und Prozess*. Master's thesis, Université de Genève.

Imperial, J. M. and Ong, E. (2021). A simple post-processing technique for improving readability assessment of texts using word mover's distance. *arXiv, http://arxiv.org/abs/2103.07277*.

Islam, M. Z. (2014). *Multilingual text classification using information-theoretic features*. Ph.D. thesis, Johann Wolfgang Goethe-Universität, Frankfurt (Main), Germany.

Karačić, J., Dondio, P., Buljan, I., Hren, D., and Marušić, A. (2019). Languages for different health information readers: multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats. *BMC medical research methodology*, **19**(1), 1–9.

Kefer, I. (2013). Die Lesbarkeit von Schulbüchern für den Betriebswirtschaftslehreunterricht. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, **109**(1), 94–107.

Keinki, C., Zowalla, R., Wiesner, M., Koester, M. J., and Huebner, J. (2018). Understandability of patient information booklets for patients with cancer. *Journal of Cancer Education*, **33**(3), 517–527.

Keinki, C., Zowalla, R., Pobiruchin, M., Huebner, J., and Wiesner, M. (2019). Computer-based readability testing of information booklets for German cancer patients. *Journal of Cancer Education*, **34**(4), 696–704.

Kercher, J. (2010). Zur Messung der Verständlichkeit deutscher Spitzenpolitiker anhand quantitativer Textmerkmale. In T. Faas, K. Arzheimer, and S. Rossteutscher, editors, *Information – Wahrnehmung – Emotion*, Schriftenreihe des Arbeitskreises „Wahlen und politische Einstellungen "der Deutschen Vereinigung für Politische Wissenschaft (DVPW), pages 97–121. VS Verlag für Sozialwissenschaften.

Kercher, J. (2011). *Verstehen und Verständlichkeit von Politikersprache: Verbale Bedeutungsvermittlung zwischen Politikern und Bürgern*. Springer VS, Hohenheim, Germany.

Kercher, J. (2013). Der Hohenheimer Komplexitätsindex für Politikersprache. In *Verstehen*

und Verständlichkeit von Politikersprache, pages 377–391. Springer.

Kercher, J. and Brettschneider, F. (2011). Nach der Wahl ist vor der Wahl? Themenschwerpunkte und Verständlichkeit der Parteien vor und nach der Bundestagswahl 2009. In *Die Parteien nach der Bundestagswahl 2009*, pages 325–353. Springer.

Kercher, J. and Brettschneider, F. (2013). Wahlprogramme als Pflichtübung? Typen, Funktionen und verständlichkeit der Bundestagswahlprogramme 1994–2009. In *Wahlen und Wähler*, pages 269–290. Springer.

Klas, K. (2011). *Der Patientenpass – ein multiprofessionelles Medium zur perioperativen Begleitung von Patientinnen und Patienten, exemplarisch am Beispiel der vorderen Kreuzbandplastik*. Master's thesis, IDS Pflegewissenschaft, Universität Wien.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing*, **23**(1), 99–130.

Lampert, A., Wien, K., Haefeli, W. E., and Seidling, H. M. (2016). Guidance on how to achieve comprehensible patient information leaflets in four steps. *International Journal for Quality in Health Care*, **28**, 634–638.

Lepper, C., Stang, J., and McElvany, N. (2021). Gender differences in text-based interest: Text characteristics as underlying variables. *Reading Research Quarterly*, **57**, 537–554.

Lin, N. and Osnabrügge, M. (2018). Making comprehensible speeches when your constituents need it. *Research & Politics*, **5**(3), 2053168018795598.

Locher, F. M., Becker, S., and Pfost, M. (2019). The relation between students' intrinsic reading motivation and book reading in recreational and school contexts. *AERA Open*, **5**(2), 233285841985204.

Luers, J. C., Gostian, A. O., Roth, K. S., and Beutner, D. (2013). Lesbarkeit von medizinischen Texten im Internetangebot deutscher HNO-Universitätskliniken. *HNO*, **61**(8), 648–654.

Lyatoshinsky, P., Pratsinis, M., Abt, D., Schmid, H. P., Zumstein, V., and Betschart, P. (2019). Readability assessment of commonly used German urological questionnaires. *Current Urology*, **13**(2), 87–93.

Merges, F. (2014). *Assistenzsystem zur Testung und Verbesserung der Lesbarkeit von Gebrauchsinformationen*. Ph.D. thesis, Naturwissenschaftlich-Technische Fakultät der Universität Siegen, Siegen, Germany.

Meyer, M. F., Bacher, R., Roth, K. S., Beutner, D., and Luers, J. C. (2013). Systematische Analyse der Lesbarkeit von Patienteninformationstexten auf Internetseiten deutscher nichtuniversitärer HNO-Kliniken. *HNO*, **62**(3), 186–195.

Muhr, R. (2012). Zur Bürgerfreundlichkeit und Verständlichkeit österreichischer Rechtstexte. *Sprachenpolitik und Rechtssprache*, pages 117–140.

Naderi, B., Mohtaj, S., Ensikat, K., and Möller, S. (2019). Automated text readability assessment for German language: a quality of experience approach. In *Eleventh international conference on quality of multimedia experience*.

Nagler, T., Lonnemann, J., Linkersdörfer, J., Hasselhorn, M., and Lindberg, S. (2014). The impact of reading material's lexical accessibility on text fading effects in children's reading performance. *Reading and Writing*, **27**(5), 841–853.

Oelke, D., Spretke, D., Stoffel, A., and Keim, D. A. (2010). Visual readability analysis: How to make your writings easier to read. In *IEEE Conference on Visual Analytics Science and Technology*, pages 123–130.

Oomen-Welke, I. (2017). Sachtexte verstehen—Dichte, Lesbarkeit, Wortschatz. In *Fachintegrierte Sprachbildung*, pages 127–150. De Gruyter Mouton.

Paul, S., Ahrend, M. D., Lüers, J. C., Roth, K. S., Grimmiger, P. P., Bopp, F., and Esfahani, B. J. (2021). Systematic analysis of readability of patient information on internet pages from departments for trauma surgery of German university hospitals. *Zeitschrift fur Orthopadie und Unfallchirurgie*, **159**(2), 187–192.

Pfeiffer, A., Kuraeva, A., Foulonneau, M., Djaghloul, Y., Tobias, E., and Ras, E. (2015). Automatically generated metrics for multilingual inline choice questions on reading comprehension. In *International Computer Assisted Assessment Conference*, pages 80–95.

Plassmann, S. and Zeidler, B. (2014). Taking decisions: Assessment for university entry. *Language Learning in Higher Education*, **4**(1), 237–255.

Plath, J. and Leiss, D. (2018). The impact of linguistic complexity on the solution of mathematical modelling tasks. *ZDM*, **50**(1), 159–171.

Radner, W., Radner, S., and Diendorfer, G. (2016). Integrating a novel concept of sentence optotypes into the RADNER reading charts. *British Journal of Ophthalmology*, **101**(3), 239–243.

Riazy, S., Simbeck, K., Traeger, M., and Woestenfeld, R. (2021). The effect of prior knowledge on persistence, participation and success in a mathematical MOOC. In *International Conference on Computer Supported Education*, pages 416–429.

Saber, M. and Weber, A. (2019). How do supermarkets and discounters communicate about sustainability? A comparative analysis of sustainability reports and in-store communication. *International Journal of Retail and Distribution Management*, **47**(11), 1181–1202.

Sander, U., Kolb, B., Christoph, C., and Emmert, M. (2016). Verständlichkeit der Texte von Qualitätsvergleichen zu Krankenhausleistungen. *Das Gesundheitswesen*, **78**(12), 828–834.

Schmitt, J. B., Schneider, F. M., Weinmann, C., and Roth, F. S. (2019). Saving tiger, orangutan & co: how subjective knowledge and text complexity influence online information seeking and behavior. *Information Communication and Society*, **22**(9), 1193–1211.

Schmitz, A. (2015). *Verständlichkeit von Sachtexten: Wirkung der globalen Textkohäsion auf das Textverständnis von Schülern*. Ph.D. thesis, Bergischen Universität Wuppertal, Institut für Bildungsforschung in der School of Education.

Schoonvelde, M., Brosius, A., Schumacher, G., and Bakker, B. N. (2019). Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS one*, **14**(2), e0208450.

Seifert, S. and Paleczek, L. (2021). Digitally assessing text comprehension in grades 3-4: Test development and validation. *Electronic Journal of e-Learning*, **19**(5), 336–348.

Soemer, A., Idsardi, H. M., Minnaert, A., and Schiefele, U. (2019). Mind wandering and reading comprehension in secondary school children. *Learning and Individual Differences*, **75**.

Spiegel, J. L., Weiss, B. G., Canis, M., and Ihler, F. (2019). Urheber, Lesbarkeit und Qualität von im Internet verfügbaren deutschsprachigen Patienteninformationen zu Hörsturz. *Laryngo-Rhino-Otologie*, **98**(S 02), 11310.

Theobald, E., Malthaner, M., and Föhl, U. (2021). KI schlägt Mensch?! Automatische Generierung von Produkttexten in Online-Shops. *HMD Praxis der Wirtschaftsinformatik*, **58**(4), 922–936.

Thoms, C., Degenhart, A., and Wohlgemuth, K. (2020). Is bad news difficult to read? A readability analysis of differently connoted passages in the annual reports of the 30 DAX companies. *Journal of Business and Technical Communication*, **34**(2), 157–187.

Tolochko, P. and Boomgaarden, H. G. (2018). Analysis of linguistic complexity in professional and citizen media. *Journalism Studies*, **19**, 1786–1803.

Tolochko, P. and Boomgaarden, H. G. (2019). Determining political text complexity: Conceptualizations, measurements, and application. *International Journal of Communication*, **13**, 21.

Toth, B. (2017). *Readability of hearing-related information on the Internet in the German language*. Master's thesis, University of Canterbury.

Vajjala, S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguis-*

tics, Processing and Educational Applications. Ph.D. thesis, Eberhard Karls Universität Tübingen, Tübingen, Germany.

Vlachos, M. and Lappas, T. (2011). Ranking German texts by comprehensibility for foreign document retrieval. In *Workshop on Enriching Information Retrieval*, Beijing, China.

vor der Brück, T. (2009). Approximation of the parameters of a readability formula by robust regression. *Intelligent Information and Communication Systems*.

Vormaier, A. (2020). *Verständlichkeit politischer Sprache in Österreich*. Master's thesis, FH Burgenland University of Applied Sciences.

Vössing, J. and Stamov-Rossnagel, C. (2016). Boosting metacomprehension accuracy in computer-supported learning: The role of judgment task and judgment scope. *Computers in Human Behavior*, **54**, 73–82.

Vössing, J., Stamov-Rossnagel, C., and Heinitz, K. (2016). Text difficulty affects metacomprehension accuracy and knowledge test performance in text learning. *Journal of Computer Assisted Learning*, **33**(3), 282–291.

Šárka Holanová (2016). Zum Charakter des gegenwärtigen Erstlesebuches in Anbetracht der Textschwierigkeit, bzw. der Lesbarkeit. Bachelor's thesis, Karlsuniversität.

Waltl, B. and Matthes, F. (2015). Comparison of law texts: an analysis of German and Austrian law texts regarding linguistic and structural metrics. In *18. Internationales Rechtsinformatik Symposium (IRIS)*, Salzburg, Austria.

Weih, M., Reinhold, A., Richter-Schmidinger, T., Sulimma, A. K., Klein, H., and Kornhuber, J. (2008). Unsuitable readability levels of patient information pertaining to dementia and related diseases: A comparative analysis. *International Psychogeriatrics*, **20**(6), 1116–1123.

Weiss, Z. (2015). More linguistically motivated features of language complexity in readability classification of german textbooks: Implementation and evaluation. Bachelor thesis. Eberhard Karls Universität Tübingen.

Weiss, Z. and Meurers, D. (2018). Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317.

Weiss, Z., Dittrich, S., and Meurers, D. (2018). A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 79–90, Stockholm, Sweden. LiU Electronic Press.

Weiss, Z., Chen, X., and Meurers, D. (2021). Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Wiesner, M., Zowalla, R., Pobiruchin, M., *et al.* (2020). The difficulty of German information booklets on psoriasis and psoriatic arthritis: automated readability and vocabulary analysis. *JMIR Dermatology*, **3**(1), e16095.

Zowalla, R., Wiesner, M., and Pfeifer, D. (2014). Automatically assessing the expert degree of online health content using SVMs. In J. Mantas, M. Househ, and A. Hasman, editors, *Integrating Information Technology and Management for Quality of Care*, volume 202 of *Studies in Health Technology and Informatics*, pages 48–51. IOS Press.

7.3 Bibliography

Abedi, J., Bayley, R., Ewers, N., Mundhenk, K., Leon, S., Kao, J., and Herman, J. (2012). Accessible reading assessments for students with disabilities. *International Journal of Disability, Development and Education*, **59**(1), 81–95.

Ács, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the sixth workshop on building and using comparable corpora*, pages 52–58.

Adelman, J. S., Brown, G. D., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science*, **17**(9), 814–823.

Adli, A., García, M. G., and Kaufmann, G. (2015). System and usage: (Never) mind the gap. In P. Auer, G. von Essen, and W. Frick, editors, *Variation in language: System-and usage-based approaches*, volume 50 of *linguae & litterae*, chapter 1, pages 1–25. De Gruyter, Berlin, Boston.

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

Akamatsu, K., Pattanasri, N., Jatowt, A., and Tanaka, K. (2011). Measuring comprehensibility of web pages based on link analysis. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 40–46. IEEE.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, **91**(4), 659–663.

Alexopoulou, T., Michel, M., Murakami, A., and Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, **67**(S1), 181–209.

Alonso, M. A., Fernandez, A., and Díez, E. (2011). Oral frequency norms for 67,979 spanish words. *Behavior Research Methods*, **43**(2), 449–458.

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service.

Andersson, K. and Szewczyk, P. (2011). Insecurity by obscurity continues: Are ADSL router manuals putting end-users at risk. In *9th Australian Information Security Management Conference*, pages 19–24.

Ashok, V. G., Feng, S., and Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, **332**(6027), 346–349.

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*, **2007**(1), i–22.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, **4**(3), 3–30.

Auguie, B. (2022). *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

Aumiller, D. and Gertz, M. (2022). Klexikon: A German dataset for joint summarization and simplification. *arXiv*.

Bachman, L. F. and Palmer, A. S., editors (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford Applied Linguistics. Oxford University Press, Oxford, United Kingdom.

Baerman, M., Brown, D., and Corbett, G. G., editors (2015). *Understanding and measuring morphological complexity*. Oxford University Press.

Bailin, A. and Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & communication*, **21**(3), 285–301.

Bamberger, R. and Vanecek, E. (1984). *Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten deutscher Sprache*. Jugend und Volk, Vienna.

Bannò, S. and Matassoni, M. (2022). Cross-corpora experiments of automatic proficiency

assessment and error detection for spoken English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 82–91, Seattle, Washington. Association for Computational Linguistics.

Bardovi-Harlig, K. (1992). A second look at t-unit analysis: Reconsidering the sentence. *TESOL quarterly*, **26**(2), 390–395.

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37**(5), 1178.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, **34**(1), 1–34.

Beers, S. F. and Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, **22**(2), 185–200.

Beinborn, L., Zesch, T., and Gurevych, I. (2012). Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, pages 11–19.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Bengoetxea, K. and Gonzalez-Dios, I. (2021). MultiAzterTest: a multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Bengoetxea, K., González-Dios, I., and Aguirregoitia, A. (2020). AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural*, **64**, 61–68.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educ Psychol Rev*, **24**, 63–88.

Bentz, C. and Berdicevskis, A. (2016). Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In *26th International Conference on Computational Linguistics (COLING 2016), 11th to 16th December 2016 Osaka, Japan.*, pages 222–232.

Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., and Trautwein, U. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, **110**, 518–543.

Berggren, S. J., Rama, T., and Øvrelid, L. (2019). Regression or classification? Automated

essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102.

Bertram, C., Weiss, Z., Zachrich, L., and Ziai, R. (2021). Artificial intelligence in history education. Linguistic content and complexity analyses of student writings in the CAHisT project (computational assessment of historical thinking). *Computers and Education: Artificial Intelligence*, page 100038.

Best, R., Ozuru, Y., Floyd, R., and McNamara, D. (2006). Children's text comprehension: Effects of genre, knowledge and text cohesion. In *Proceedings of the 7th international conference on Learning sciences*, pages 37–42. International Society of the Learning Sciences.

Best, R. M., Floyd, R. G., and McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading psychology*, **29**(2), 137–164.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, **69**, 65–78.

Bhat, S. and Yoon, S.-Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, **67**, 42–57.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, pages 384–414.

Biber, D. and Conrad, S. (2001). Register variation: A corpus approach. In D. Schiffrin, D. Tannen, and H. E. Hamilton, editors, *The handbook of discourse analysis*, pages 175–196. Blackwell Publishers Ltd.

Biber, D., Gray, B., and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, **45**(1), 5–35.

Biber, D., Gray, B., and Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, **37**(5), 639–668.

Biberauer, T., Holmberg, A., Roberts, I., and Sheehan, M. (2014). Complexity in comparative syntax: The view from modern parametric theory. In F. J. Newmeyer and L. B. Preston, editors, *Measuring Grammatical Complexity*, Oxford Linguistics, chapter 6, pages 103–127. Oxford University Press, Oxford, United Kingdom, 1st edition.

Bingel, J., Paetzold, G., and Søgaard, A. (2018). Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational*

Linguistics, pages 245–258.

Birdsong, D. and Gertken, L. M. (2013). In faint praise of folly: A critical review of native/non-native speaker comparisons, with examples from native and bilingual processing of French complex syntax. *Language, Interaction and Acquisition*, **4**(2), 107–133.

Björnsson, C.-H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, pages 480–497.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, **33**(1), 1–17.

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Bonfiglio, T. P. (2010). *Mother tongues and nations*. De Gruyter Mouton.

Bonin, P., Barry, C., Méot, A., and Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and language*, **50**(4), 456–476.

Borade, J. G. and Netak, L. D. (2020). Automated grading of essays: a review. In *International Conference on Intelligent Human Computer Interaction*, pages 238–249. Springer.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Brants, T., Skut, W., and Uszkoreit, H. (1999). Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA Treebank Workshop*, Paris.

Breindl, E., Volodina, A., and Waßner, U. H. (2014). *Handbuch der deutschen Konnektoren 2*, volume 13 of *Schriften des Instituts für Deutsche Sprache*. Walter de Gruyter GmbH & Co KG.

Brezina, V. and Pallotti, G. (2019). Morphological complexity in written l2 texts special issue on linguistic complexity. *Second Language Research*, **35**, 99–119.

Briscoe, T., Medlock, B., and Andersen, Ø. (2010). Automated assessment of ESOL free text examinations. Technical report, University of Cambridge, Computer Laboratory.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, **40**(2), 540–545.

Brown, J. and Eskenazi, M. (2004). Retrieval of authentic documents for reader-specific lexi-

cal practice. In *InSTIL/ICALL Symposium 2004*.

Brück, T. V. D. and Hartrumpf, S. (2007a). A readability checker based on deep semantic indicators. In *Human Language Technology. Challenges of the Information Society: Third Language and Technology Conference, LTC 2007, Poznan, Poland, October 5-7, 2007, Revised Selected Papers 3*.

Brück, T. V. D. and Hartrumpf, S. (2007b). A semantically oriented readability checker for German. In *Proceedings of the 3rd Language and Technology Conference*, pages 270–274.

Brück, T. V. D. and Leveling, J. (2007). Parameter learning for a readability checking tool. In *LWA*.

Brück, T. V. D., Helbig, H., and Leveling, J. (2008). The readability checker delite technical report. Technical report.

Brunato, D., Dell'Orletta, F., Venturi, G., François, T., and Blache, P., editors (2016). *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, Osaka, Japan.

Brunet, É. (1978). *Le vocabulaire de Jean Giraudoux structure et évolution. Statistique et informatique appliquées à l'étude des textes, à partir du Trésor de la langue française*. Le vocabulaire des grands écrivains français. Slatkine, Genève.

Bryant, L., Spencer, E., Ferguson, A., Craig, H., Colyvas, K., and Worrall, L. (2013). Propositional Idea Density in aphasic discourse. *Aphasiology*, **27**(8), 992–1009.

Brysbaert, M. and Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology*, **64**(3), 545–559.

Brysbaert, M. and Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual cognition*, **13**(7-8), 992–1011.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, **41**(4), 977–990.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., and Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental psychology*, **58**(5), 412.

Brysbaert, M., Lagrou, E., and Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, **20**(3), 530–548.

Brysbaert, M., Mandera, P., and Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, **27**(1), 45–50.

Budd, R., Marius, T., Gatewood, P., and Jones, D. (2019). Using k-means in SVR-based text difficulty estimation. In *Workshop on Speech and Language Technology in Education*, pages 84–88. International Speech Communication Association.

Bulté, B. (2013). *The development of complexity in second language acquisition. A dynamic systems approach*. Ph.D. thesis, Vrije Universiteit Brussel, Brussels.

Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, I. Vedder, and F. Kuiken, editors, *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, volume 32 of *Language Learning & Language Teaching*, chapter 2, pages 23–46. John Benjamins Publishing, Amsterdam/Philadelphia.

Bulté, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of second language writing*, **26**, 42–65.

Bulté, B. and Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, **28**(1), 147–164.

Bulté, B. and Housen, A. (2019). Beginning L2 complexity development in CLIL and non-CLIL secondary education. *Instructed Second Language Acquisition*, **3**(2), 153–180.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, **25**, 60–117.

Burstein, J. and Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Computer mediated language assessment and evaluation in natural language processing*, pages 68–75.

Cacoullos, R. T. and Travis, C. E. (2019). Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics*, **57**(3), 653–692.

Cain, K. (2003). Text comprehension and its relation to coherence and cohesion in children's fictional narratives. *British Journal of Developmental Psychology*, **21**(3), 335–351.

Caines, A. and Buttery, P. (2020). REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5614–5623.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, **3**(1), 1–27.

Cameron, C. A., Lee, K., Webster, S., Munro, K., Hunt, A. K., and Linton, M. J. (1995). Text cohesion in children's narrative writing. *Applied Psycholinguistics*, **16**(3), 257–269.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards and R. Schmidt, editors, *Language and communication*, pages 2–27. Routledge, London.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, **1**(1), 1–47.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Chen, J., Fife, J. H., Bejar, I. I., and Rupp, A. A. (2016). Building e-rater® scoring models using machine learning methods. *ETS Research Report Series*, **1**, 1–12.

Chen, X. (2018). *Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research*. Ph.D. thesis, Eberhard Karls Universität Tübingen Germany.

Chen, X. and Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 113–119, Osaka, Japan. COLING.

Chen, X. and Meurers, D. (2018). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, **41**(3), 486–510.

Chen, X. and Meurers, D. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, **32**, 418–447.

Chen, Y.-Y., Liu, C.-L., Lee, C.-H., Chang, T.-H., *et al.* (2010). An unsupervised automated essay-scoring system. *IEEE Intelligent systems*, **25**(5), 61–67.

Chiari, I. and De Mauro, T. (2014). The new basic vocabulary of Italian as a linguistic resource. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014*, pages 113–116, Pisa, Italy. Pisa University Press.

Chinkina, M. and Meurers, D. (2016). Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*, pages 188–198.

Chinkina, M., Kannan, M., and Meurers, D. (2016). Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 7–12, Berlin, Germany. Association for Computational Linguistics. `http://anthology.aclweb.org/`

P16-4002.

Chodorow, M. and Burstein, J. (2004). Beyond essay length: evaluating e-rater®'s performance on TOEFL® essays. *ETS Research Report Series*, **2004**(1), i–38.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, **2**(3), 113–124.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, **70**(4), 213.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, **60**(2), 283.

Collins-Thompson, K. (2014). Computational assessment of text readability. a survey of current and future research. *ITL - International Journal of Applied Linguistics*, **165**, 97–135.

Collins-Thompson, K. and Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, pages 193–200.

Collins-Thompson, K., Bennett, P. N., White, R. W., De La Chica, S., and Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412.

Constantin, A.-E. and Patil, I. (2021). ggsignif: R package for displaying significance brackets for 'ggplot2'. *PsyArxiv*.

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., and Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, **55**(3), 493–520.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press/Council of Europe, Cambridge, UK.

Cox, B. E., Shanahan, T., and Sulzby, E. (1990). Good and poor elementary readers' use of cohesion in writing. *Reading research quarterly*, **25**(1), 47–65.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, **34**(2), 213–238.

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, **11**, 415–443.

Crossley, S. and McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, pages 984–989.

Crossley, S. and McNamara, D. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, pages 1236–1241. Cognitive Science Society.

Crossley, S. and McNamara, D. (2016a). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Grantee Submission*, **7**, 351–370.

Crossley, S., Salsbury, T., and McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, **59**(2), 307–334.

Crossley, S., Salsbury, T., and McNamara, D. (2010a). The development of polysemy and frequency use in English second language speakers. *Language Learning*, **60**(3), 573–605.

Crossley, S., Allen, L. K., Snow, E. L., and McNamara, D. S. (2015). Pssst... textual features... there is more to automatic essay scoring than just you! In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 203–207.

Crossley, S., Russell, D., Kyle, K., and Romer, U. (2017). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics*, **1**, 48–81.

Crossley, S. A. and McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, **18**(2), 119–135.

Crossley, S. A. and McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, **35**(2), 115–135.

Crossley, S. A. and McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, **26**, 66–79.

Crossley, S. A. and McNamara, D. S. (2016b). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, **7**(3), 351–370.

Crossley, S. A., Greenfield, J., and McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, **42**(3), 475–493.

Crossley, S. A., Salsbury, T., and McNamara, D. S. (2010b). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, **29**(2), 243–263.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., and McNamara, D. S. (2011a). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, **28**(3), 282–311.

Crossley, S. A., Salsbury, T., McNamara, D. S., and Jarvis, S. (2011b). What is lexical proficiency? Some answers from computational models of speech data. *Tesol Quarterly*, **45**(1), 182–193.

Crossley, S. A., Salsbury, T., and McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, **29**(2), 243–263.

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., and McNamara, D. S. (2014a). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Grantee Submission*, **7**(1).

Crossley, S. A., Roscoe, R., and McNamara, D. S. (2014b). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, **31**(2), 184–214.

Crossley, S. A., Yang, H. S., and McNamara, D. S. (2014c). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, **26**(1), 92–113.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, **32**, 1–16.

Crossley, S. A., Allen, L. K., Snow, E. L., and McNamara, D. S. (2016b). Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining*, **8**(2), 1–19.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016c). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, **48**(4), 1227–1237.

Crossley, S. A., Skalicky, S., and Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, **42**(3-4), 541–561.

Cuetos, F., Alvarez, B., González-Nosti, M., Méot, A., and Bonin, P. (2006). Determinants of lexical access in speech production: Role of word frequency and age of acquisition. *Memory & cognition*, **34**(5), 999–1010.

Cuetos, F., Rodríguez-Ferreiro, J., Sage, K., and Ellis, A. W. (2012a). A fresh look at the predictors of naming accuracy and errors in Alzheimer's disease. *Journal of neuropsychology*, **6**(2), 242–256.

Cuetos, F., Glez-Nosti, M., Barbón, A., and Brysbaert, M. (2012b). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, **33**(2), 133–143.

Cui, Y., Zhu, J., Yang, L., Fang, X., Chen, X., Wang, Y., and Yang, E. (2022). CTAP for Chinese: a linguistic complexity feature automatic calculation platform. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5525–5538.

Cummins, J. (1997). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, **19**, 198–205.

Cummins, J. (2000). BICS and CALP. *Encyclopedia of language teaching and learning*, pages 76–79.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street and N. H. Hornberger, editors, *Encyclopedia of language and education*, volume 2, pages 71–83. Springer Science + Business Media LLC, New York, 2nd edition.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*, volume 10. John Benjamins Amsterdam.

Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*, **26**, 19–26.

Daller, H., Hout, R. V., and Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, **24**, 197–222.

Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., and Trausan-Matu, S. (2018). Please ReaderBench this text: A multi-dimensional textual complexity assessment framework. In *Tutoring and intelligent tutoring systems*, pages 251–271. Nova Science Publishers, Inc.

De, A. and Kopparapu, S. K. (2011). An unsupervised approach to automated selection of good essays. In *2011 IEEE Recent Advances in Intelligent Computational Systems*, pages 662–666. IEEE.

De Clercq, B. and Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, **101**(2), 315–334.

De Clercq, B. and Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research. Special Issue*

on Linguistic Complexity, **35**, 71–97.

De Clercq, O. and Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, **42**(3), 457–490.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., and Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, **20**(3), 293–325.

De Groot, C. (2008). Morphological complexity as a parameter of linguistic typology. *Language complexity: Typology, contact, change*, page 191.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, **54**(2), 113–132.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., and Hulstijn, J. H. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, I. Vedder, and F. Kuiken, editors, *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, volume 32 of *Language Learning & Language Teaching*, chapter 6, pages 121–142. John Benjamins Amsterdam, Amsterdam/Philadelphia.

De Meo, A., Maffia, M., and Vitale, G. (2019). La competenza scritta in italiano L2 di apprendenti vulnerabili. Due scale di valutazione a confronto. *EL. LE*, **8**(3), 637–654.

Deane-Mayer, Z. A. and Knowles, J. E. (2019). *caretEnsemble: Ensembles of Caret Models*.

Dehrmann, M.-G. (2014). Die Austreibung der Schrift durch die Schrift. Zur philologisch-historischen Reflexion von Mündlichkeit nach 1800 am Beispiel der Grimmschen Kinder- und Hausmärchen. *Fabula*, **55**(1-2), 153–170.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, **109**(2), 193–210.

Deng, Z., Peng, H., Xia, C., Li, J., He, L., and Yu, P. (2020). Hierarchical bi-directional self-attention networks for paper review rating recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6302–6314, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Deutsch, T., Jasbi, M., and Shieber, S. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA. Association for Computational Linguistics.

Deygers, B. (2021). The CEFR companion volume: Between research-based policy and policy-based research. *Applied Linguistics*, **42**(1), 186–191.

Diependaele, K., Lemhöfer, K., and Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, **66**(5), 843–863.

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., and Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in psychology*, **1**, 218.

Dittrich, S., Weiss, Z., Schröter, H., and Meurers, D. (2019). Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education. *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*.

Doerr, N. M. (2009). Investigating "native speaker effects": Toward a new model of analyzing "native speaker" ideologies. In N. M. Doerr, editor, *The Native Speaker Concept Ethnographic Investigations of Native Speaker Effects*, number 26 in Language, Power and Social Process, chapter 1, pages 15–46. Mouton de Gruyter Berlin.

Dong, F. and Zhang, Y. (2016). Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1072–1077.

Dowle, M. and Srinivasan, A. (2021). *data.table: Extension of 'data.frame'*.

Druskat, S., Gast, V., Krause, T., and Zipser, F. (2016). Corpus-tools.org: An interoperable generic software tool set for multi-layer linguistic corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4492–4499.

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

DuBay, W. H. (2006). The classic readability studies. *Online Submission*.

Durrant, P. (2016). To what extent is the academic vocabulary list relevant to university student writing? *English for specific purposes*, **43**, 49–61.

Dussias, P. E. (2001). Psycholinguistic complexity in codeswitching. *International Journal of Bilingualism*, **5**(1), 87–100.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, **41**(1), 87–100.

Ehret, K. (2018). Kolmogorov complexity as a universal measure of language complexity. *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 8–14.

Ehret, K. and Szmrecsanyi, B. (2016). An informationtheoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, **57**, 71–94.

Ehret, K. and Szmrecsanyi, B. (2019). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research. Special Issue on Linguistic Complexity*, **35**(1), 23–45.

Eisenberg, P., Peters, J., Gallmann, P., Fabricius-Hansen, C., Nübling, D., Barz, I., and Fiehler, R. (2009). *Duden. Deutsche Grammatik*, volume 4. Bibliographisches Institut & F.A. Brockhaus AG, Mannheim, Germany, 8 edition.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, **24**(2), 143–188.

Ellis, N. C. and Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—introduction to the special issue. *Applied linguistics*, **27**(4), 558–589.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford university press.

Englert, C. S. and Hiebert, E. H. (1984). Children's developing awareness of text structures in expository materials. *Journal of educational psychology*, **76**(1), 65–74.

Eraslan, S., Yaneva, V., Yesilada, Y., and Harper, S. (2017). Do web users with autism experience barriers when searching for information within web pages? In *Proceedings of the 14th International Web for All Conference*, pages 1–4.

Eraslan, S., Yesilada, Y., Yaneva, V., and Ha, L. A. (2021). "Keep it simple!": An eye-tracking study for exploring complexity and distinguishability of web pages for people with autism. *Universal Access in the Information Society*, **20**(1), 69–84.

Esfandiari, R. and Ahmadi, M. (2022). Phraseological complexity and academic writing proficiency in abstracts authored by student and expert writers. *English Teaching & Learning*, pages 1–20.

Esses, V. M. and Maio, G. R. (2002). Expanding the assessment of attitude components and structure: The benefits of open-ended measures. *European review of social psychology*, **12**(1), 71–101.

Fausey, C. M. and Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-

witness memory. *Psychonomic bulletin & review*, **18**(1), 150–157.

Fedorenko, E., Woodbury, R., and Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive science*, **37**(2), 378–394.

Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, I. Vedder, and F. Kuiken, editors, *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, volume 32 of *Language Learning & Language Teaching*, pages 277–297. John Benjamins Amsterdam, Netherlands, Amsterdam/Philadelphia.

Ferwerda, J., Hainmueller, J., and Hazlett, C. J. (2017). Kernel-based regularized least squares in R (KRLS) and Stata (krls). *Journal of Statistical Software*, **79**(3), 1–26.

Fitzgerald, J. and Spiegel, D. L. (1986). Textual cohesion and coherence in children's writing. *Research in the Teaching of English*, **20**(3), 263–280.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, **32**, 221.

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, chapter 9, pages 167–184. Psychology Press, New York.

Foster, P. and Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language learning*, **59**(4), 866–896.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics*, **21**(3), 354–375.

François, T. and Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.

Frey, J.-C. (2020a). Age-related language in South Tyrolean social media. In *Using data mining to repurpose German language corpora. An evaluation of data-driven analysis methods*

for corpus linguistics, pages 198–250. Università di Bologna.

Frey, J.-C. (2020b). *Using data mining to repurpose German language corpora. An evaluation of data-driven analysis methods for corpus linguistics*. Ph.D. thesis, Università di Bologna.

Freyhoff, G., Hess, G., Kerr, L., Menzell, E., Tronbacke, B., and Van Der Veken, K. (1998). *Make It Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability for authors, editors, information providers, translators and other interested persons*. International League of Societies for Persons with Mental Handicap European Association, Brussels.

Futrell, R. and Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics*, volume 1 (long papers), pages 688–698, Valencia, Spain. Association for Computational Linguistics.

Futrell, R., Gibson, E., and Levy, R. P. (2021). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, **44**(3), e12814.

Galasso, S. (2014). Exploring textual cohesion characteristics for German readability classification. In *BA thesis. Eberhard Karls Universität Tübingen*.

Gardner, D. and Davies, M. (2014). A new academic vocabulary list. *Applied linguistics*, **35**(3), 305–327.

Gardner, S., Nesi, H., and Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, **40**(4), 646–674.

Garner, J., Crossley, S., and Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, **80**, 176–187.

Gee, J. (1990). *Social linguistics and literacies: Ideologies in Discourses*. Falmer Press, New York.

Gerndt, H. (1988). Sagen und Sagenforschung im Spannungsfeld von Mündlichkeit und Schriftlichkeit. *Fabula*, **29**, 1–20.

Gibbons, P. (1991). *Learning to learn in a second language*. Primary English Teaching Association, Newton, Australia.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. M. Alec P. Marantz and W. O'Neil, editors, *Image, Language, Brain. Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press.

Gierut, J. A. (2007). Phonological complexity and language learnability. *American Journal of Speech-Language Pathology*, **16**(1), 6–17.

Gimenes, M. and New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior research methods*, **48**(3), 963–972.

Glöckner, I., Hartrumpf, S., Helbig, H., Leveling, J., and Osswald, R. (2006). An architecture for rating and controlling text readability. In *Proceedings of KONVENS*, pages 32–35.

Gnewuch, U., Morana, S., Heckmann, C., and Maedche, A. (2018). Designing conversational agents for energy feedback. In *International Conference on Design Science Research in Information Systems and Technology*, pages 18–33. Springer.

Gobrecht, B. (1997). Schweizerdeutsche Märchen zwischen Mündlichkeit und Schriftlichkeit. *Fabula*, **38**(1-2), 42–64.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. *Proceedings of the 8th International Language Ressources and Evaluation*, pages 759–765.

Gonzalez-Garduno, A. and Søgaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Göpferich, S. and Neumann, I. (2016). Writing competence profiles as an assessment grid? – Students' L1 and L2 writing competences and their development after one semester of instruction. In *Developing and Assessing Academic and Professional Writing Skills*, pages 103–140. Peter Lang, Bern, Switzerland.

Graesser, A. C., McNamara, D. S., and Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking reading comprehension*, **82**, 98.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, **36**, 193–202.

Gramacy, R. B., Moler, C., and Turlach, B. A. (2022). *monomvn: Estimation for MVN and Student-t Data with Monotone Missingness*.

Greenfield, J. (1999). *Classic readability formulas in an EFL context: Are they valid for*

Japanese speakers? Ph.D. thesis, Temple Univesity.

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal*, **26**(1), 5–24.

Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, **29**(2), 261–290.

Guerini, M., Pepe, A., and Lepri, B. (2012). Do linguistic style and readability of scientific abstracts affect their virality? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 475–478.

Gülzow, I. and Gagarina, N. V. (2007). Noun phrases, pronouns and anaphoric reference in young children narratives. *ZAS papers in linguistics*, **48**, 203–223.

Guo, L., Crossley, S. A., and McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, **18**(3), 218–238.

Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., and Hervas-Martinez, C. (2015). Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, **28**(1), 127–146.

Gyllstrom, K. and Moens, M.-F. (2010). Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 159–168.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Hamp, B. and Feldweg, H. (1997). GermaNet – a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Hancke, J. (2013). *Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language*. Master's thesis, Eberhard Karls Universität Tübingen, Tübingen, Germany.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic and morphological features. In *Proceedings of COLING*, pages 1063–1080.

Hansen-Schirra, S. and Maaß, C. (2020). Easy language, plain language, easy language plus:

Perspectives on comprehensibility and stigmatisation. In S. Hansen-Schirra and C. Maaß, editors, *Easy Language Research: Text and User Perspectives*, volume 2 of *Easy – Plain – Accessible*, chapter 2, pages 17–38. Frank & Thieme, Berlin, Germany.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., and Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86. Association for Computational Linguistics.

Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 295–304.

Hawlik, R. and Sorger, B. (2017). DBK: eine Sprachstandserhebung über den Aneignungsstand schriftlicher Bildungssprache in der Primar-und Sekundarstufe I. *Open Online Journal for Research and Education*.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics (proceedings of the main conference)*, pages 460–467.

Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M., Juffs, A., and Wilson, L. (2010). Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, **20**(1), 73–98.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., and Kliegl, R. (2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, **62**, 10–20.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, **21**(248), 1–43.

Hennig, M. (2017). *Linguistische Komplexität—ein Phantom?*, chapter 1, pages 7–18. Stauffenburg Verlag.

Hennig, M. and Niemann, R. (2013). Unpersönliches Schreiben in der Wissenschaft: Eine Bestandsaufnahme. *Informationen Deutsch als Fremdsprache*, **40**(4), 439–460.

Hills, T. T., Maouene, J., Riordan, B., and Smith, L. B. (2010). The associative structure of

language: Contextual diversity in early word learning. *Journal of memory and language*, **63**(3), 259–273.

Hirao, R., Arai, M., Shimanaka, H., Katsumata, S., and Komachi, M. (2020). Automated essay scoring system for nonnative Japanese learners. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1250–1257.

Hoffman, P., Lambon Ralph, M. A., and Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, **45**(3), 718–730.

Honkela, T., Izzatdust, Z., and Lagus, K. (2012). Text mining for wellbeing: Selecting stories using semantic and pragmatic features. In *International Conference on Artificial Neural Networks*, pages 467–474. Springer.

Hornung, R. (2021). *ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables*.

Horst, M. and Collins, L. (2006). From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review*, **63**(1), 83–106.

Höttecke, D., Feser, M. S., Heine, L., and Ehmke, T. (2018). Do linguistic features influence item difficulty in physics assessments? *Science Education Review Letters*, pages 1–6.

Housen, A. and Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, **30**(4), 461–473.

Housen, A., Vedder, I., and Kuiken, F. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, volume 32 of *Language Learning & Language Teaching*. John Benjamins Amsterdam, Amsterdam/Philadelphia.

Housen, A., Clercq, B. D., Kuiken, F., and Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research. Special Issue on Linguistic Complexity*, **35**(1), 3–21.

Howcroft, D. M. and Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

Hsiao, Y. and Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, **103**, 114–126.

Hu, R., Wu, J., and Lu, X. (2022). Word-combination-based measures of phraseological diversity, sophistication, and complexity and their relationship to second language chinese

proficiency and writing quality. *Language Learning*.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, **91**(4), 663–667.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, **8**(3), 229–249.

Hulstijn, J. H. (2014). The common european framework of reference for languages: A challenge for applied linguistics. *ITL-International Journal of Applied Linguistics*, **165**(1), 3–18.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers*. John Benjamins Publishing, Amsterdam.

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., and Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common european framework of reference for languages (CEFR). *Language Testing*, **29**(2), 203–221.

Hunt, K. W. (1965). A synopsis of clause-to-sentence length factors. *The English Journal*, **54**(4), 300–309.

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *Tesol Quarterly*, pages 195–202.

Hussein, M. A., Hassan, H., and Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, **5**, e208.

Imperial, J. M. (2021). BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618.

Imperial, J. M. and Ong, E. (2021). A simple post-processing technique for improving readability assessment of texts using word mover's distance. *arXiv*.

Jackson, D. O. and Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, **63**(2), 330–367.

Jameel, S. and Qian, X. (2012). An unsupervised technical readability ranking model by building a conceptual terrain in LSI. In *2012 Eighth International Conference on Semantics, Knowledge and Grids*, pages 39–46. IEEE.

Jameel, S., Qian, X., and Lam, W. (2012). N-gram fragment sequence based unsupervised

domain-specific document readability. In *Proceedings of COLING 2012*, pages 1309–1326.

Jarrold, W. L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H. S., and Swan, G. E. (2010). Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease. In *International Conference on Brain Informatics*, pages 299–307. Springer.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, **63**, 87–106.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, **34**(4), 537–553.

Jin, W. (2001). A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency levels. In *Paper presented at the Symposium on Second Language Writing at Purdue University*.

Johns, B. T. and Jones, M. N. (2022). Content matters: Measures of contextual diversity must consider semantic content. *Journal of Memory and Language*, **123**, 104313.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., and Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, **132**(2), EL74–EL80.

Johns, B. T., Dye, M., and Jones, M. N. (2016). The influence of contextual variability in word learning. *Psychonomic Bulletin & Review*, **23**, 1214–1220.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, **37**, 13–38.

Joseph, H. S., Bremner, G., Liversedge, S. P., and Nation, K. (2015). Working memory, reading ability and the effects of distance and typicality on anaphor resolution in children. *Journal of Cognitive Psychology*, **27**(5), 622–639.

Jung, Y., Crossley, S., and McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *Journal of Asia TEFL*, **16**(1), 37–52.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, and F. Karlsson, editors, *Language complexity: Typology, contact, change*, volume 94, pages 89–108. John Benjamins Publishing, Amsterdam, Philadelphia.

Kaggle (2012). The hewlett foundation: Automated essay scoring. Develop an automated scoring algorithm for student-written essays. `https://www.kaggle.com/competitions/asap-aes/overview/`.

Kamalski, J., Sanders, T., and Lentz, L. (2008). Coherence marking, prior knowledge, and

comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, **45**(4-5), 323–345.

Karatzoglou, A., Smola, A., and Hornik, K. (2022). *kernlab: Kernel-Based Machine Learning Lab*.

Karlsson, F., Miestamo, M., and Sinnemäki, K. (2008). The problem of language complexity. In M. Miestamo, K. Sinnemäki, and F. Karlsson, editors, *Language complexity: Typology, contact, change*, volume 94, pages vii–xiv. John Benjamins Publishing, Amsterdam, Philadelphia.

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*.

Kassim, N. L. A. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, **11**(3), 179–197.

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

Ke, Z. and Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 6300–6308.

Kim, J. Y., Collins-Thompson, K., Bennett, P. N., and Dumais, S. T. (2012). Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222.

Kim, M., Crossley, S. A., and Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, **102**(1), 120–141.

King, M. M., Winton, A. S., and Adkins, A. D. (2003). Assessing the readability of mental health internet brochures for children and adolescents. *Journal of child and family studies*, **12**(1), 91–99.

Kintsch, W. (1974). *The representation of meaning in memory*. Erlbaum, Hillsdale, NJ.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, **95**(2), 163.

Kintsch, W. and Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive psychology*, **5**(3), 257–274.

Kiwanuka, E., Mehrzad, R., Prsic, A., and Kwan, D. (2017). Online patient resources for

gender affirmation surgery: An analysis of readability. *Annals of Plastic Surgery*, **79**, 329–333.

Kormos, J. and Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, **62**(2), 439–472.

Kortmann, B. and Szmrecsanyi, B. (2012). Introduction: Linguistic complexity Second Language Acquisition, indigenization, contact. In B. Kortmann and B. Szmrecsanyi, editors, *Linguistic complexity: Second language acquisition, indigenization, contact*, linguae & litterae, pages 6–34. de Gruyter, Walter GmbH & Co, Berlin, Boston.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman, New York.

Krause, T. and Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, **31**(1), 118–139.

Kühberger, C., Bramann, C., Weiss, Z., and Meurers, D. (2019). Task complexity in history textbooks: A multidisciplinary case study on triangulation in history education research. *History Education Research Journal*, **16**.

Kuhn, M. (2022). *caret: Classification and Regression Training*.

Kuiken, F. and Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken, and I. Vedder, editors, *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, pages 143–170. Benjamins Philadelphia, PA.

Kuiken, F. and Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, **34**(3), 321–336.

Kuiken, F. and Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. *Task-based approaches to teaching and assessing pragmatics*, pages 266–285.

Kuiken, F. and Vedder, I. (2022). Measurement of functional adequacy in different learning contexts. In F. Kuiken and I. Vedder, editors, *TASK Journal on Task-Based Language Teaching and Learning*, volume 2, chapter 2, pages 8–32. John Benjamins Publishing.

Kuiken, F., Vedder, I., and Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, **1**, 81–100.

Kuiken, F., Vedder, I., Housen, A., and Clercq, B. D. (2019). Variation in syntactic complexity: Introduction. *International Journal of Applied Linguistics (United Kingdom)*, **29**, 161–170.

Kusters, W. (2008). Complexity in linguistic theory, language learning and language change.

In *Language complexity: Typology, contact, change*, volume 94, pages 3–22. John Benjamins Publishing, Amsterdam, Philadelphia.

Kyle, K. and Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, **34**, 12–24.

Kyle, K. and Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, **34**(4), 513–535.

Kyle, K. and Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, **102**(2), 333–349.

Kyle, K., Crossley, S. A., and Jarvis, S. (2021a). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, **18**(2), 154–170.

Kyle, K., Crossley, S., and Verspoor, M. (2021b). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, **43**(4), 781–812.

Laarmann-Quante, R., Ortmann, K., Ehlert, A., Masloch, S., Scholz, D., Belke, E., and Dipper, S. (2019). The Litkey corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods*, **51**(4), 1889–1918.

Lambert, C. and Nakamura, S. (2019). Proficiency-related variation in syntactic complexity: A study of English L1 and L2 oral descriptive discourse. *International Journal of Applied Linguistics*, **29**(2), 248–264.

Langevin, R., Lordon, R., and Avrahami, T. (2021). Heuristic evaluation of conversational agents. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in Second Language Acquisition. *Applied Linguistics*, **30**(4), 579–589.

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, **16**(3), 307–322.

Lavalley, R., Berkling, K., and Stüker, S. (2015). Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

Lee, B. W. and Lee, J. (2020). LXPER Index 2.0: Improving text readability assessment model for L2 English students in Korea. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–24.

Leng, Y., Yu, L., and Xiong, J. (2019). Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review. In *2019 International Conference on Multi-*

modal Interaction, pages 395–403.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning*, **40**(3), 387–417.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel, editor, *Sentence processing*, pages 90–126. Psychology Press.

Levy, R. and Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Li, M. and Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*, volume 3 of *Texts in Computer Science*. Springer, 4th edition.

Li, Y. and Qian, D. D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System*, **38**(3), 402–411.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, **2**(3), 18–22.

Lin, J., Song, J., Zhou, Z., and Shi, X. (2021). Automated scholarly paper review: Possibility and challenges. *arXiv preprint arXiv:2111.07533*.

Lipson, M. Y. (1982). Learning new information from text: The role of prior knowledge and reading ability. *Journal of reading behavior*, **14**(3), 243–261.

Liu, J. and Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, **39**, 1–11.

Long, D. L., Johns, C. L., and Morris, P. E. (2006). Comprehension ability in mature readers. In M. J. Traxler and M. A. Gernsbacher, editors, *Handbook of Psycholinguistics*, chapter 20, pages 801–833. Elsevier, 2nd edition edition.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., and Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, pages 843–848.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, **15**(4), 474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL quarterly*, **45**(1), 36–62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, **96**(2), 190–208.

Lüdeling, A. and Kytö, M. (2008). *Corpus linguistics: An international handbook*, volume 29

of *Handbücher zur Sprach- und Kommunikationswissenschaft*. Walter de Gruyter GmbH, Berlin, New York.

Ludewig, U., Trendtel, M., Weiss, Z., Meurers, D., and McElvany, N. (2022). What text features make reading comprehension difficult across elementary school? Investigating difficulty and changes in difficulty. *PsyArXiv*.

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., and Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, **3**, 897–915.

Lumley, T. (2020). *leaps: Regression Subset Selection*.

Lupyan, G. and Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, **5**(1), e8559.

Maddieson, I. (2009). Calculating phonological complexity. In F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé, editors, *Approaches to phonological complexity*, volume 16 of *Phonology and Phonetics*, pages 85–110. Walter de Gruyter, Berlin / New York.

Maddieson, I., Bhattacharya, T., Smith, D. E., and Croft, W. (2011). Geographical distribution of phonological complexity. *Linguistic Typology*, **15**, 267–279.

Madrazo Azpiazu, I. and Pera, M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, **7**, 421–436.

Madrazo Azpiazu, I. and Pera, M. S. (2020a). An analysis of transfer learning methods for multilingual readability assessment. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–100.

Madrazo Azpiazu, I. and Pera, M. S. (2020b). Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*.

Malvern, D., Richards, B., Chipere, N., and Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Springer.

Maniyar, U., Joseph, K. J., Deshmukh, A. A., Dogan, U., and Balasubramanian, V. N. (2020). Zero shot domain generalization. In *Proceedigs of the 31st British Machine Vision Virtual Conference*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. `http://aclweb.org/anthology/P/P14/P14-5010`.

Marchisio, K., Guo, J., Lai, C.-I., and Koehn, P. (2019). Controlling the reading level of

machine translation output. *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203.

Marcus, S., Kerz, E., Wiechmann, D., and Neumann, S. (2016). Cocogen-complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window technique. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 23–31.

Martin, M., Mustonen, S., Reiman, N., and Seilonen, M. (2010). On becoming an independent user. *Communicative Profiency and Linguistic Development: Intersections between SLA and Language Testing Research. Eurosla Monographs*, **1**, 57–79.

Martinc, M., Pollak, S., and Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, **47**(1), 141–179.

Marujo, L., Lopes, J., Mamede, N., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J., and Viana, C. (2009). Porting REAP to European Portuguese. In *International Workshop on Speech and Language Technology in Education*.

McCannon, B. C. (2019). Readability and research impact. *Economics Letters*, **180**, 76–79.

McCarthy, K. S. and McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist*, **56**(3), 196–214.

McCarthy, K. S., Guerrero, T. A., Kent, K. M., Allen, L. K., McNamara, D. S., Chao, S.-F., Steinberg, J., O'Reilly, T., and Sabatini, J. (2018). Comprehension in a scenario-based assessment: Domain and topic-specific background knowledge. *Discourse Processes*, **55**(5-6), 510–524.

McCarthy, P. and Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment: An ecological validation study. In S. Jarvis and M. Daller, editors, *Vocabulary Knowledge: Human ratings and automated measures*, pages 45–78. John Benjamins, Amsterdam / Philadelphia.

McCarthy, P. M. and Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, **24**(4), 459–488.

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, **42**(2), 381–392.

McClellan, C. A. (2010). Constructed-response scoring—doing it right. *R & D Connections*, **13**, 1–7.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of

text sequence and prior knowledge. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **55**(1), 51.

McNamara, D. S. and Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy and C. Boonthum, editors, *Applied natural language processing: Identification, investigation and resolution*, pages 188–205. IGI Global, Hershey, PA.

McNamara, D. S. and Kintsch, W. (1996). Learning from texts: effects of prior knowledge and text coherence. *Discourse Processes*, **22**, 247–288.

McNamara, D. S., Kintsch, E., and Songer, N. B. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Kintsch Source: Cognition and Instruction*, **14**, 1–43.

McNamara, D. S., Cai, Z., and Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, chapter 19, pages 391–412. Psychology Press, New York.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010a). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, **47**(4), 292–330.

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010b). Linguistic features of writing quality. *Written communication*, **27**(1), 57–86.

McNamara, D. S., Ozuru, Y., and Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International electronic journal of elementary education*, **4**(1), 229–257.

McWhorter, J. (2008). Why does a language undress? Strange cases in Indonesia. In M. Miestamo, K. Sinnemäki, and F. Karlsson, editors, *Language complexity: Typology, contact, change*, volume 94, pages 167–190. John Benjamins Publishing, Amsterdam, Philadelphia.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typology*, **5**, 125–156.

Mecklenburg, N. (2018). Vom Sagennachklang zum Gesellschaftsecho. Spuren von Mündlichkeit in erzählten Gesprächen. In *Theodor Fontane. Realismus, Redevielfalt, Ressentiment*, pages 76–89. J. B. Metzler.

Meng, C., Chen, M., Mao, J., and Neville, J. (2020). Readnet: A hierarchical transformer framework for web article readability analysis. In *European Conference on Information Retrieval*, pages 33–49. Springer.

Menn, L. and Duffield, C. J. (2014). Looking for a 'Gold Standard' to measure language

complexity: What psycholinguistics and neurolinguistics can (and cannot) offer to formal linguistics. In F. J. Newmeyer and L. B. Preston, editors, *Measuring Grammatical Complexity*, Oxford Linguistics, chapter 6, pages 103–127. Oxford University Press, Oxford, United Kingdom, 1st edition.

Messick, S. (1996). Validity and washback in language testing. *Language testing*, **13**(3), 241–256.

Metallinou, A. and Cheng, J. (2014). Syllable and language model based features for detecting non-scorable tests in spoken language proficiency assessment applications. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–98.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2022). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*. Technical University of Vienna.

Microsoft Corporation and Weston, S. (2022). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*.

Miestamo, M. (2004). On the feasibility of complexity metrics. In K. Kerge and M.-M. Sepper, editors, *Finest Linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, pages 11–26, Tallinn. Publications of the Department of Estonian of Tallinn University 8.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In *Language complexity: Typology, contact, change*, volume 94, pages 23–42. John Benjamins Publishing, Amsterdam, Philadelphia.

Miestamo, M., Sinnemäki, K., and Karlsson, F. (2008). *Language complexity: Typology, contact, change*, volume 94 of *Studies in Language Companion Series*. John Benjamins Publishing, Amsterdam, Philadelphia.

Miller, R. T., Mitchell, T. D., and Pessoa, S. (2016). Impact of source texts and prompts on students' genre uptake. *Journal of Second Language Writing*, **31**, 11–24.

Miltsakaki, E. (2009). Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 49–52.

Miltsakaki, E. and Troutt, A. (2008). Real time web text classification and analysis of reading difficulty. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 89–97.

Misra, P., Agarwal, N., Kasabwala, K., Hansberry, D. R., Setzen, M., and Eloy, J. A. (2013).

Readability analysis of healthcare-oriented education resources from the american academy of facial plastic and reconstructive surgery. *The Laryngoscope*, **123**(1), 90–96.

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford university press.

Mohammadi, H. and Khasteh, S. H. (2019). Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv:1912.05957*.

Mohammadi, H. and Khasteh, S. H. (2020). A machine learning approach to Persian text readability assessment using a crowdsourced dataset. In *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, pages 1–7. IEEE.

Monaghan, W. and Bridgeman, B. (2005). E-rater® as a quality control on human scores. *R & D Connections*.

Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Monteiro, R. A., Santos, R. L., Pardo, T. A., Almeida, T. A. d., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. In *International Conference on Computational Processing of the Portuguese Language*, pages 324–334. Springer.

Mueller, S. T., Seymour, T. L., Kieras, D. E., and Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **29**(6), 1353.

Myford, C. M. and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, **4**(4), 386–422.

Myhill, D. (2008). Towards a linguistic model of sentence development in writing. *Language and Education*, **22**(5), 271–288.

Naderi, B., Mohtaj, S., Ensikat, K., and Möller, S. (2019a). Automated text readability assessment for German language: a quality of experience approach. In *Eleventh international conference on quality of multimedia experience*.

Naderi, B., Mohtaj, S., Ensikat, K., and Möller, S. (2019b). Subjective assessment of text complexity: A dataset for German language. *arXiv*.

Neri, N. C. and Klückmann, F. (2021). LATIC–a linguistic analyzer for text and item characteristics documentation (version 1.1. 0). `https://www.researchgate.net/`

publication/351579029_LATIC_-_A_Linguistic_Analyzer_for_Text_and_Item_Characteristics_Documentation_Version_110.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**(3), 516–524.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., and Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, **6**, 312.

Nomura, M., Nielsen, G. S., and Tronbacke, B. (2010). Guidelines for easy-to-read materials. revision on behalf of the ifla/library services to people with special needs section. IFLA Professional Reports 120, International Federation of Library Associations and Institutions, The Hague, IFLA Headquarters.

Norris, J. and Ortega, L. (2003). Defining and measuring SLA. *The handbook of second language acquisition*, pages 716–761.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics*, **30**(4), 555–578.

Oelke, D., Spretke, D., Stoffel, A., and Keim, D. A. (2010). Visual readability analysis: How to make your writings easier to read. In *IEEE Conference on Visual Analytics Science and Technology*, pages 123–130.

Oh, B.-D., Clark, C., and Schuler, W. (2021). Surprisal estimators for human reading times need character models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1 (long papers). Association for Computational Linguistics.

Ojha, P. K., Ismail, A., and Kuppusamy, K. (2018). Readability assessment-cum-evaluation of government department websites of Rajasthan. In *Proceedings of First International Conference on Smart System, Innovations and Computing*, pages 235–244. Springer.

Okinina, N., Jennifer-Carmen, F., and Weiss, Z. (2020). CTAP for Italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*.

Olinghouse, N. G. and Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second-and fourth-grade students. *Reading and Writing*, **22**(5), 545–565.

Olinghouse, N. G. and Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, **26**(1), 45–65.

O'Reilly, T. and McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes*, **43**(2), 121–152.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, **24**(4), 492–518.

Ortega, L. (2012). Interlanguage complexity. In B. Kortmann and B. Szmrecsanyi, editors, *Linguistic complexity: Second language acquisition, indigenization, contact*, volume 13 of *linguae & litterae*, pages 127–155. Walter de Gruyter, Berlin, Boston.

Ortega, L. and Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual review of applied linguistics*, **25**, 26–45.

Östling, R., Smolentzov, A., Hinnerich, B. T., and Höglin, E. (2013). Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

Ott, N. and Meurers, D. (2011). Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education*, **3**(1-2), 9–30.

Ott, N. and Ziai, R. (2010). Evaluating dependency parsing performance on German learner language. In M. Dickinson, K. Müürisep, and M. Passarotti, editors, *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, volume 9 of *NEALT Proceeding Series*, pages 175–186, Tartu, Estonia. Tartu University Press.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In T. Schmidt and K. Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Ozuru, Y., Dempsey, K., and McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and instruction*, **19**(3), 228–242.

Padó, U. (2016). Get semantic with me! The usefulness of different feature types for short-answer grading. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers*, pages 2186–2195.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, **47**(5), 238–243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International review of education*, pages 210–225.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., *et al.* (2021). The PRISMA 2020

statement: an updated guideline for reporting systematic reviews. *Systematic reviews*, **10**(1), 1–11.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, **30**, 590–601.

Pallotti, G. (2015a). A simple view of linguistic complexity. *Second Language Research*, **31**, 117–134.

Pallotti, G. (2015b). Una nuova misura della complessità linguistica: l'Indice di Complessità Morfologica (ICM). [A new measure of linguistic complexity: the Morphological Complexity Index (MCI)]. *Rivista Italiana di Linguistica Applicata (RILA)*, **2**(3), 195–215.

Pallotti, G. (2017). Applying the interlanguage approach to language teaching. *International review of applied linguistics in language teaching*, **55**(4), 393–412.

Pallotti, G. (2019). An approach to assessing the linguistic difficulty of tasks. *Journal of the European Second Language Association*, **3**, 58–70.

Pallotti, G. and Ferrari, S. (2008). La variabilità situazionale dell'interlingua: implicazioni per la ricerca acquisizionale e il testing linguistico. *Competenze lessicali e discorsive nell'acquisizione di lingue seconde*, pages 437–461.

Pancer, E., Chandler, V., Poole, M., and Noseworthy, T. J. (2019). How readability shapes social media engagement. *Journal of Consumer Psychology*, **29**(2), 262–270.

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, **15**(1), 29–43.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second language research*, **35**(1), 121–145.

Paul, S., Ahrend, M. D., Lüers, J. C., Roth, K. S., Grimmiger, P. P., Bopp, F., and Esfahani, B. J. (2021). Systematic analysis of readability of patient information on internet pages from departments for trauma surgery of german university hospitals. *Zeitschrift fur Orthopadie und Unfallchirurgie*, **159**, 187–192.

Pellegrino, F., Marsico, E., Chitoran, I., and Coupé, C., editors (2009). *Approaches to phonological complexity*, volume 16 of *Phonology and Phonetics*. Walter de Gruyter, Berlin / New York.

Peña, E. D., Bedore, L. M., and Torres, J. (2021). Assessment of language proficiency and dominance in monolinguals and bilinguals. In W. S. Francis, editor, *Bilingualism Across the Lifespan: Opportunities and Challenges for Cognitive Research in a Global Society*,

Frontiers in Cognitive Psychology, chapter 6, pages 88–105. Routledge.

Pera, M. S. and Ng, Y.-K. (2012). Brek12: A book recommender for K-12 users. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1037–1038.

Perrig, W. and Kintsch, W. (1985). Propositional and situational representations of text. *Journal of Memory and language*, **24**(5), 503–518.

Peterson, B. G. and Carl, P. (2020). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York.

Phandi, P., Chai, K. M. A., and Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439. Asso.

Piggin, G. (2012). What are our tools really made out of? A critical assessment of recent models of language proficiency. *Polyglossia: the Asia-Pacific's voice in language and language teaching*, **22**, 79–87.

Pilán, I., Vajjala, S., and Volodina, E. (2016). A readable read: automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, **7**(1), 143–159.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Plakans, L. and Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, **22**(3), 217–230.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report.

Polenz, P. v. (1981). Über die Jargonisierung von Wissenschaftssprache und wider die Deagentivierung. In T. Bungarten, editor, *Wissenschaftliche Beiträge zur Methodologie, theoretischen Fundierung und Deskription*, pages 85–110. Fink, München.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., and Kukich, K. (2002). Stump-

ing e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, **18**(2), 103–134.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Quinlan, T., Higgins, D., and Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series*, (1), 1–35.

Quispesaravia, A., Perez, W., Cabezudo, M. S., and Alva-Manchego, F. (2016). Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4694–4698.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahimi, Z. and Litman, D. (2016). Automatically extracting topical components for a response-to-text writing assessment. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 277–282, San Diego, CA. Association for Computational Linguistics.

Rahimi, Z., Litman, D. J., Correnti, R., Matsumura, L. C., Wang, E., and Kisa, Z. (2014). Automatic scoring of an analytical response-to-text assessment. In *International conference on intelligent tutoring systems*, pages 601–610. Springer.

Ramineni, C. and Williamson, D. (2018). Understanding mean score differences between the e-rater automated scoring engine and humans for demographically based groups in the GRE general test. *ETS Research Report Series*, **2018**(1), 1–31.

Reali, F., Chater, N., and Christiansen, M. H. (2014). The paradox of linguistic complexity and community size. In *Evolution of language: Proceedings of the 10th international conference (evolang10)*, pages 270–277. World Scientific.

Rello, L., Saggion, H., Baeza-Yates, R., and Graells, E. (2012). Graphical schemes may improve readability but not understandability for people with dyslexia. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 25–32.

Rello, L., Baeza-Yates, R., Dempere-Marco, L., and Saggion, H. (2013a). Frequent words improve readability and short words improve understandability for people with dyslexia. In

IFIP Conference on Human-Computer Interaction, pages 203–219. Springer.

Rello, L., Bautista, S., Baeza-Yates, R., Gervás, P., Hervás, R., and Saggion, H. (2013b). One half or 50%? An eye-tracking study of number representation readability. In *IFIP Conference on Human-Computer Interaction*, pages 229–245. Springer.

Rello, L., Baeza-Yates, R., Bott, S., and Saggion, H. (2013c). Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Rescher, N. (1998). *Complexity: A philosophical overview*. Transaction Publishers, London.

Resnik, P. (1992). Left-corner parsing and psychological plausibility. In *The 14th International Conference on Computational Linguistics*, pages 191–197, Nantes, France. Association for Computational Linguistics.

Révész, A., Ekiert, M., and Torgersen, E. N. (2014). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, **37**(6), 828–848.

Reynolds, R. (2016). Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA, USA. Association for Computational Linguistics.

Riebling, L. (2013). Heuristik der bildungssprache. *Herausforderung Bildungssprache–und wie man sie meistert*, pages 106–153.

Riemenschneider, A., Weiss, Z., Schröter, P., and Meurers, D. (2021). Linguistic complexity in teachers' assessment of German essays in high stakes testing. *Assessing Writing*, **50**.

Rigutini, L. and Algherini, S. (2022). Towards sustainable technology: "green" approaches to NLP. A comparative study in terms of performance and energy consumption. `https://towardsdatascience.com/toward-sustainable-technology-green-approaches-to-nlp-81a9fb2cf521`. Last checked: November, 17, 2022.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language learning*, **45**(1), 99–140.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, **22**(1), 27–57.

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the

SSARC model of pedagogic task sequencing. In M. Bygate, editor, *Domains and directions in the development of TBLT. A decade of plenaries from the international conference*, volume 8 of *Task-based language teaching*, chapter 5, pages 87–122. John Benjamins Publishing, Amsterdam.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., and McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, **34**, 39–59.

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA.

Rubin, R. (2021). Assessing the impact of automatic dependency annotation on the measurement of phraseological complexity in L2 Dutch. *International Journal of Learner Corpus Research*, **7**(1), 131–162.

Russell, D. (2011). Searchresearch. Search by reading level. `https://searchresearch1.blogspot.com/2011/02/search-by-reading-level.html`. Last checked: October, 7th, 2022.

Saddiki, H., Habash, N., Cavalli-Sforza, V., and Al-Khalil, M. (2018). Feature optimization for predicting readability of Arabic L1 and L2. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29.

Samuda, V. and Bygate, M. (2008). Defining pedagogic tasks: Issues and challenges. In *Tasks in Second Language Learning*, pages 62–70. Palgrave Macmillan UK, London.

Santhanam, S., Karduni, A., and Shaikh, S. (2020). Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Santos, R., Pedro, G., Leal, S., Vale, O., Pardo, T., Bontcheva, K., and Scarton, C. (2020). Measuring the impact of readability features in fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1404–1413, Marseille, France. European Language Resources Association.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York.

Scarton, C. and Aluısio, S. M. (2010). Coh-Metrix-Port: a readability assessment tool for texts in Brazilian Portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*, volume 10. sn.

Schmidt, K. (2016). Der graphematische Satz [The graphematic sentence]. *Zeitschrift für*

Germanistische Linguistik, **44**(2), 215–256.

Schriver, K. A. (2000). The mechanism used by readability formulas makes them unreliable. Readability formulas in the new millennium: What's the use? *ACM Journal of Computer Documentation (JCD)*, **24**(3), 138–140.

Schuwirth, L. W. and Van Der Vleuten, C. P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical education*, **38**(9), 974–979.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Seeker, W. and Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a German treebank. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation, 3132–3139. Istanbul, Turkey: European Language Resources Association (ELRA)*.

Seidenspinner, W. (1997). Oralisierte Schriftlichkeit als Stil. Das literarische Genre Dorfgeschichte und die Kategorie Mündlichkeit. *Internationales Archiv für Sozialgeschichte der Deutschen Literatur*, **22**(2), 36–51.

Seifried, E., Lenhard, W., and Spinath, B. (2016). Automatic essay assessment: Effects on students' acceptance and on learning-related characteristics. *psihologija*, **49**(4), 469–482.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, **10**, 209–232.

Shadrova, A., Linscheid, P., Lukassek, J., Lüdeling, A., and Schneider, S. (2021). A challenge for contrastive L1/L2 corpus studies: Large inter- and intra-individual variation across morphological, but not global syntactic categories in task-based corpus data of a homogeneous L1 German group. *Frontiers in Psychology*, **12**, 5267.

Shain, C., Schijndel, M. V., Futrell, R., Gibson, E., and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 49–58.

Shannon, C. (1948). A mathematical theory of communication. *The Bell systems technical journal*, **27**, 379–424, 623–656.

Shelley, M. C. and Schuh, J. H. (2001). Are the best higher education journals really the best? A meta-analysis of writing quality and readability. *Journal of scholarly Publishing*, **33**(1), 11–22.

Shen, A., Qi, J., and Baldwin, T. (2017). A hybrid model for quality assessment of Wikipedia

articles. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 43–52.

Shen, A., Salehi, B., Baldwin, T., and Qi, J. (2019). A joint model for multimodal document quality assessment. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 107–110.

Shen, W., Williams, J., Marius, T., and Salesky, E. (2013). A language-independent approach to automatic text difficulty assessment for second-language learners. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, **165**(2), 259–298.

Siegel, J. (2010). *Second dialect acquisition*. Cambridge University Press, Cambridge.

Singh, A. D., Mehta, P., Husain, S., and Rajkumar, R. (2016). Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 202–212.

Sinnemäki, K. (2020). Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics*, **6**(2), 20191010.

Sirts, K., Piguet, O., and Johnson, M. (2017). Idea density for predicting Alzheimer's disease from transcribed speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 322–332.

Sitbon, L. and Bellot, P. (2008). A readability measure for an information retrieval process adapted to dyslexics. In *Second international workshop on Adaptive Information Retrieval (AIR 2008 in conjunction with IIiX 2008)*, pages 52–57.

Skehan, P. (1998). *A Cognitive Approach to Learning Language*. Oxford University Press.

Skehan, P. (2009). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller, editors, *Vocabulary studies in first and second language acquisition. The Interface Between Theory and Application*, pages 107–124. Palgrave Macmillan London.

Skehan, P. (2015). Limited attention capacity and cognition. In M. Bygate, editor, *Domains and directions in the development of TBLT. A decade of plenaries from the international*

conference, volume 8 of *Task-based language teaching*, chapter 6, pages 123–155. John Benjamins Publishing, Amsterdam.

Skierkowski, D. D., Florin, P., Harlow, L. L., Machan, J., and Ye, Y. (2019). A readability analysis of online mental health resources. *American Psychologist*, **74**(4), 474.

Smith, M., Breakstone, J., and Wineburg, S. (2019). History assessments of thinking: A validity study. *Cognition and Instruction*, **37**(1), 118–144.

Smith, R., Snow, P., Serry, T., and Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, **42**(3), 214–240.

Spiegel, D. L. and Fitzgerald, J. (1990). Textual cohesion and coherence in children's writing revisited. *Research in the Teaching of English*, **24**(1), 48–66.

Stahns, R. (2016). Bildungssprachliche Merkmale von Texten und Items. Zur Operationalisierung des Konstrukts „Bildungssprache ". *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, **21**(41), 44–55.

Štajner, S. (2021). Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What can readability measures really tell us about text complexity? In L. Rello and H. Saggion, editors, *In Proceedings of the First Workshop on Natural Language Processing for Improving Textual Accessibility*. European Language Resources Association (ELRA).

Staples, S., Egbert, J., Biber, D., and Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, **33**(2), 149–183.

Stevenson, M. and Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, **19**, 51–65.

Ströbel, M., Kerz, E., and Wiechmann, D. (2020). The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning*, **70**(3), 732–767.

Struthers, L., Lapadat, J. C., and MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing*, **18**(3), 187–201.

Stymne, S., Tiedemann, J., Hardmeier, C., and Nivre, J. (2013). Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 378–474.

Sung, Y.-T., Lin, W.-C., Dyson, S. B., Chang, K.-E., and Chen, Y.-C. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, **99**(2), 371–391.

Sunyaev, A., Dehling, T., Taylor, P. L., and Mandl, K. D. (2015). Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association*, **22**, e28–e33.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.

Taguchi, N., Crawford, W., and Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *Tesol Quarterly*, **47**(2), 420–430.

Tan, C., Gabrilovich, E., and Pang, B. (2012). To each his own: Personalized content selection based on text comprehensibility. In *In Proceedings of WSDM*.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, **58**(1), 1–21.

Tavakoli, P. and Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, **61**, 37–72.

Tavakoli, P. and Skehan, P. (2005). Strategic planning, task structure, and performance testing. *Planning and task performance in a second language*, **239273**.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lissabon.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language learning*, **44**, 307–307.

Thorndike, E. L. (1921). *The teacher's word book*. Teacher's College, Columbia University.

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A.-L., and Bernhard, D. (2013). Coherence and cohesion for the assessment of text readability. In *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013)*, pages 11–19.

Todirascu, A., François, T., Bernhard, D., Gala, N., and Ligozat, A.-L. (2016). Are cohesive features relevant for text readability evaluation? In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 987–997.

Tracy-Ventura, N. and Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, **1**(1),

58–95.

Trotzke, A. and Zwart, J.-W. (2014). The complexity of narrow syntax: Minimalism, representational economy, and simplest merge. In F. J. Newmeyer and L. B. Preston, editors, *Measuring Grammatical Complexity*, Oxford Linguistics, chapter 7, pages 128–147. Oxford University Press, Oxford, United Kingdom, 1st edition.

Turner, A. and Greene, E. (1977). The construction and use of a propositional text base. Technical Report 63, Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, **48**(2), 459–484.

Vajjala, S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, Eberhard Karls Universität Tübingen, Tübingen, Germany.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, **28**(1), 79–105.

Vajjala, S. (2022). Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Vajjala, S. and Lõo, K. (2013). Role of morpho-syntactic features in Estonian proficiency classification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72.

Vajjala, S. and Lõo, K. (2014). Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127.

Vajjala, S. and Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Vajjala, S. and Lučić, I. (2019). On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop*

on building educational applications using NLP, pages 163–173. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2014). Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*.

Vajjala, S., Meurers, D., Eitel, A., and Scheiter, K. (2016). Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48.

Valencia, S. W., Wixson, K. K., and Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, **115**(2), 270–289.

van der Slik, F., Hout, R. v., and Schepens, J. (2019). The role of morphological complexity in predicting the learnability of an additional language: The case of La (additional language) Dutch. *Second Language Research*, **35**(1), 47–70.

van Schijndel, M., Nguyen, L., and Schuler, W. (2013). An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 37–46, Sofia, Bulgaria. Association for Computational Linguistics.

Vandeweerd, N., Housen, A., and Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study. *International Journal of Learner Corpus Research*, **7**(2), 197–229.

VanPatten, B. and Benati, A. G. (2010). *Key terms in second language acquisition*. Bloomsbury Publishing.

Vargas, C. R., Ricci, J. A., Lee, M., Tobias, A. M., Medalie, D. A., and Lee, B. T. (2017). The accessibility, readability, and quality of online resources for gender affirming surgery. *Journal of Surgical Research*, **217**, 198–206.

Vasylets, O., Gilabert, R., and Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, **67**(2), 394–430.

Venant, R. and d'Aquin, M. (2019). Towards the prediction of semantic complexity based on concept graphs. In *12th International Conference on Educational Data Mining (EDM 2019)*, pages 188–197.

Vercellotti, M. L. (2019). Finding variation: assessing the development of syntactic complex-

ity in ESL speech. *International Journal of Applied Linguistics*, **29**(2), 233–247.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied psycholinguistics*, **22**(2), 217–234.

Verspoor, M., Schmid, M. S., and Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, **21**(3), 239–263.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., and Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, **39**, 50–63.

vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica*, **32**(4).

Vyatkina, N., Hirschmann, H., and Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*, **29**, 28–50.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Wahlen, A., Kuhn, C., Zlatkin-Troitschanskaia, O., Gold, C., Zesch, T., and Horbach, A. (2020). Automated scoring of teachers' pedagogical content knowledge – a comparison between human and machine scoring. *Frontiers in Education*, **5**.

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., and Quintana, R. (2020). eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, **44**, 100449.

Wang, J., Liang, S.-l., and Ge, G.-c. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, **27**(4), 442–458.

Wang, Z. and von Davier, A. A. (2014). Monitoring of scoring using the e-rater® automated scoring system and human raters on a writing test. *ETS Research Report Series*, (1), 1–21.

Watson, R. and Kochmar, E. (2021). Read & Improve: A novel reading tutoring system. In *EDM 2021: Educational Data Mining*.

Weiss, Z. (2015). More linguistically motivated features of language complexity in readability classification of German textbooks: Implementation and evaluation. Bachelor Thesis in Computational Linguistics.

Weiss, Z. (2017). Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects. In *BA thesis. Eberhard Karls Universität Tübingen*.

Weiss, Z. and Meurers, D. (2018). Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317.

Weiss, Z. and Meurers, D. (2019a). Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school. In *Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School*, pages 380–393.

Weiss, Z. and Meurers, D. (2019b). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In A. Abel, A. Glaznieks, V. Lyding, and L. Nicolas, editors, *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference. Corpora and Language in Use – Proceedings 5*, pages 419–435. Presses universitaires de Louvain.

Weiss, Z. and Meurers, D. (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task. Using proficiency classification to investigate short answer data, cross-data generalizablity, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, **7**, 83–130.

Weiss, Z. and Meurers, D. (2022). Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *17th Workshop on Innovative Use of NLP for Building Educational Applications*.

Weiss, Z., Dittrich, S., and Meurers, D. (2018). A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*.

Weiss, Z., Riemenschneider, A., Schröter, P., and Meurers, D. (2019). Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 30–45.

Weiss, Z., Chen, X., and Meurers, D. (2021). Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Weiss, Z., Lange-Schubert, K., Geist, B., and Meurers, D. (2022). Sprachliche Komplexität im Unterricht: Eine computerlinguistische Analyse der gesprochenen Sprache von Lehrenden und Lernenden im naturwissenschaftlichen Unterricht in der Primar- und Sekundarstufe.

Zeitschrift für Germanistische Linguistik.

Wendt, D., Brand, T., and Kollmeier, B. (2014). An eye-tracking paradigm for analyzing the processing time of sentences with different linguistic complexities. *PLoS One*, **9**(6), e100186.

Whalen, T. E. (1971). The analysis of essays by computer: A simulation of teacher' ratings. In *Paper presented at the Annual leeting of the American Educational Research Association*, New York. ERIC.

Wharton, C. and Kintsch, W. (1991). An overview of construction-integration model: a theory of comprehension as a foundation for a new cognitive architecture. *ACM Sigart Bulletin*, **2**(4), 169–173.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wickham, H. and Bryan, J. (2022). *readxl: Read Excel Files*.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**, 1686.

Wisniewski, K. (2011). The empirical validity of the CEFR fluency scale: the A2 level description. In *Exploring language frameworks: Proceedings of the ALTE Kraków conference*, pages 253–272.

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning*, **67**(S1), 232–253.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel, A., and Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013*.

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity. Technical report, Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Manoa, Hawaii.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, **77**(1), 1–17.

Xanthos, A. and Gillis, S. (2010). Quantifying the development of inflectional diversity. *First language*, **30**(2), 175–198.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., Gagarina, N., Hrzica, G., Ketrez, F. N., Kilani-Schoch, M., Korecky-Kröll, K., Kovac ˇevic´, M., Laalo, K., Palmovic´, M., Pfeiler, B., Voeikova, M. D., and Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, **31**(4), 461–479.

Xia, M., Kochmar, E., and Briscoe, T. (2016). Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22. Association for Computational Linguistics.

Xiao, Y. and Watson, M. (2017). Guidance on conducting a systematic literature review. *https://doi.org/10.1177/0739456X17723971*, **39**, 93–112.

Xie, S., Evanini, K., and Zechner, K. (2012). Exploring content features for automated speech scoring. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 103–111.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, **3**, 283–297.

Yan, Y. (2016). *MLmetrics: Machine Learning Evaluation Metrics*.

Yaneva, V. (2015). Easy-read documents as a gold standard for evaluation of text simplification output. In *Proceedings of the Student Research Workshop*, pages 30–36.

Yaneva, V., Temnikova, I., and Mitkov, R. (2015). Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 49–57.

Yaneva, V., Temnikova, I., and Mitkov, R. (2016). Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 293–299.

Yang, W., Lu, X., and Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, **28**, 53–67.

Yannakoudakis, H. and Cummins, R. (2015). Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th annual meeting of the association*

for computational linguistics: human language technologies, pages 180–189. Association for Computational Linguistics.

Yin, K. M. (1985). The role of prior knowledge in reading comprehension. *Reading in a Foreign Language*, **3**(1), 375–380.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, **66**, 130–141.

Yoon, H.-J. (2018). The development of ESL writing quality and lexical proficiency: Suggestions for assessing writing achievement. *Language Assessment Quarterly*, **15**(4), 387–405.

Yoon, H.-J. and Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *Tesol Quarterly*, **51**(2), 275–301.

Yuill, N. and Oakhill, J. (1988). Understanding of anaphoric relations in skilled and less skilled comprehenders. *British Journal of Psychology*, **79**(2), 173–186.

Zakaluk, B. L. and Samuels, S. J. (1988). *Readability: its past, present, and future*. International Reading Association.

Zechner, K., Higgins, D., Xi, X., and Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, **51**(10), 883–895.

Zesch, T. and Horbach, A. (2018). ESCRITO – an NLP-enhanced educational scoring toolkit. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232. Association for Computational Linguistics.

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L., Howe, E., and Quintana, R. (2019). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9619–9625.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, **21**(2), 1–11.

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*, **56**, 40–53.

Ziai, R. (2018). *Short Answer Assessment in Context: The Role of Information Structure*.

Ph.D. thesis, Eberhard Karls Universität Tübingen.

Zipser, F. and Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*.

Zwaan, R. A. and Rapp, D. N. (2006). Discourse comprehension. In M. J. Traxler and M. A. Gernsbacher, editors, *Handbook of Psycholinguistics*, chapter 18, pages 725–764. Elsevier, 2nd edition edition.

Chapter 8

Publications

The publications and manuscripts listed in the previous section are enclosed in the following pages. All articles are sorted by order of appearance in Chapter 5.

Weiss, Z. & Meurers, D. (2019). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In Andrea Abel, Aivars Glaznieks, Verena Lyding & Lionel Nicolas (eds) *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Corpora and Language in Use – Proceedings 5, Louvain-la-Neuve: Presses universitaires de Louvain, 419-435.

Broad linguistic modeling is beneficial for German L2 proficiency assessment

Zarah Weiss & Detmar Meurers

University of Tübingen[1]

Abstract

We investigate the applicability of a broad range of language features to German second language proficiency assessment by comparing the performance of classification models based on linguistically diverse vs. homogeneous feature groups in terms of their overall performance and their success at individual proficiency levels (A1 to C1/C2). For this, we extract 400 measures of linguistic complexity from the domains of syntax, lexicon, morphology, discourse, language use, and human language processing. Overall, our results show that a broad feature set integrating aspects of language as a system, language use, and human sentence processing costs results in higher classification performance on language learner data. At individual proficiency levels, lexical complexity in particular, but also clausal and phrasal complexities as well as discourse measures successfully distinguish several proficiency levels. Morphological complexity is particularly important for more advanced learners.

1. Introduction

This study investigates the applicability of a broad range of language features to German second language (L2) proficiency assessment. We focus on aspects

1 http://icall-research.de

of *Complexity*, a core component of the triad *Complexity, Accuracy, and Fluency* (CAF) that is used in Second Language Acquisition (SLA) research to characterize language performance (Housen *et al.* 2012). In recent years, diverse features have been proposed to measure language proficiency, readability, and writing skills (Bulté & Housen 2014; Ortega 2012). They differ in terms of the nature of the language characteristics they measure, their specificity, sensitivity to task-effects, and how difficult it is to extract the information. To which extent the combination of diverse features is beneficial, as far as we are aware, has not been systematically investigated. Furthermore, while it is common to introduce the benefits of so far under-researched domains of linguistic complexity, such as morphology or SLA based features, a detailed comparison of which domains of linguistic complexity discriminate best at certain levels of proficiency has less often been attempted. We address this by comparing (1) performance differences between German L2 proficiency classifiers based on either broad, linguistically diverse or homogeneous feature groups; and (2) performance differences of linguistically homogeneous classifiers at individual proficiency levels (A1 to C1/C2). We find that linguistically diverse proficiency models that combine features from various linguistic domains systematically outperform those informed by individual linguistic domains. Regarding the informativeness of these linguistic domains, we find that single features from all linguistic domains that we measured are highly informative. Regarding the contribution of feature groups comprised from one linguistic domain to the identification of individual proficiency levels, we find in particular lexical, but also clausal and phrasal complexities as well as discourse measures to be highly successful across levels.

The remainder of the article is structured as follows. We briefly review previous work on the link between linguistic complexity and different levels of L2 proficiency, before outlining our automatic complexity analysis approach. We then introduce the data from the Merlin corpus which we use for our analyses. This is followed by our classification study and an outlook on current and future work, before we conclude with some final remarks.

2. Related work

Automatic complexity analyses for proficiency assessment often focus on longitudinal English L2 data elicited in University contexts for a certain group of L2 learners, such as intermediate or advanced learners, rather than distinguishing multiple proficiency levels at once. Thus, they focus more on developmental patterns that may be observed within a group of learners.

Comparisons of proficiency levels are mostly based on the collective evaluation of multiple studies targeting different learner groups, which may be potentially problematic due to different study set-ups or operationalizations of complexity.

Overall, such studies have found that learners at low and intermediate proficiency levels predominantly develop in terms of sentence length and clausal elaborateness (Lu 2010; Norris & Ortega 2009; Ortega 2003), the latter of which has been shown to be correlated with human proficiency ratings (Crossley & McNamara 2014). For more advanced English L2 learners, research indicates a stronger development of the phrasal domain, in particular regarding noun phrases (Crossley & McNamara 2014; Taguchi *et al.* 2013). Other complexity measures, such as lexical and clausal complexity, were found to be less informative to distinguish between advanced-intermediate and advanced learners (Paquot 2017; Ortega 2012; Biber *et al.* 2011). Some studies indicate that advanced learners also develop in terms of lexical abstractness, lexical familiarity, and semantic inter-relatedness (Crossley *et al.* 2014; Crossley & McNamara 2012), but that this development is not necessarily considered for advanced proficiency ratings (Crossley & McNamara 2014). As for discourse measures, studies for more advanced learners have found that more proficient learners use more implicit cohesion markers and less explicit markers, such as connectives (Crossley *et al.* 2014; Crossley & McNamara 2012; McNamara *et al.* 2009).

While there is an extensive body of research on English L2 development, there is overall less research on German complexity assessment, most of which focuses on German readability assessment (Hancke *et al.* 2012; vor der Brück *et al.* 2008). Hancke & Meurers (2013) investigate how measures of clausal, lexical, and morphological complexity as well as language model features relate to CEFR ratings. They find lexical and morphological complexities to be most informative, and clausal complexity, while less informative on its own, to boost classification performance when being combined with the other measures, and that a combined model of features overall performs best with accuracy values of 62.7%. In this study, we follow up on these results on the same corpus with a broader set of features and analytical methods.

3. Automatic complexity analysis

For our automatic analysis of German linguistic complexity, the elaborateness and variation in the different domains of linguistic modeling (Ellis & Barkhuizen 2005), we extracted 400 features using an elaborate NLP tool chain.

3.1. Complexity measures

Our features may broadly be grouped into two categories: those targeting dimensions of the theoretical linguistic system (syntax, lexicon, morphology, and discourse) and those targeting the cognitive or psycholinguistic dimension of language productions (language use and human language processing). Additionally, we calculate two descriptive, superficial features of text length in words and sentences.

Clausal and phrasal complexities assess syntactic complexity development on two levels: clausal complexity is associated with phenomena such as clausal subordination and the use of syndetic and asyndetic constructions. Our system measures in particular various types of subordination and clausal structure (t-units per sentence, dependent clauses per t-unit, etc.). Phrasal complexity measures aspects of phrasal modification and coordination. We assess these in terms of various modifier ratios and coverage of modifier measures with a particular focus on the nominal domain and verb clusters.

Lexico-semantic complexity is typically associated with vocabulary range (lexical density and variation) and size (lexical sophistication), but also lexical relatedness. We measure lexical diversity using raw type token ratio as well as its length normalized variants. Lexical density and variation are assessed for various Parts-of-Speech (PoS), including for example verb and noun variations. To measure lexical relatedness, we assess the number of semantic relations between words (hyponymy, synonymy, etc.) using GermaNet 9.0.1 (Henrich & Hinrichs 2010).

Morphological complexity has shown to be particularly interesting for languages that exhibit richer morphology than English, such as German or French (François & Fairon 2012; Hancke *et al.* 2012). We measure features of inflection (tenses, person, number, etc.), derivation (in particular nominalizations), and composition.

Discourse measures assess textual cohesion, i.e. the linguistic items that link propositions or idea units, which has been shown to be highly informative in previous work on complexity assessment among others by the *CohMetrix* project. Following them, we measure co-referential cohesion in terms of noun, argument, stem, and content-word overlaps, pronoun ratios, and various types of connectives. We also adopted transitional features from Barzilay & Lapata (2008) that assess changes in grammatical functions (subject, object, other complement, not present) that are assigned to repeated linguistic material in adjacent sentences.

Language use measures origin from psycho- and corpus-linguistic research and include, for example, word frequencies from representative language samples or approximations of age of acquisition. Word frequency measures are well established in complexity research, although they are often listed among features of lexical complexity. We calculate a series of word frequency measures based on frequency data bases Subtlex-DE and Google Books 2000-2009 (Brysbaert *et al.* 2011), as well as dlexDB (Heister *et al.* 2011). These features include absolute and log transformed frequencies. We also approximate average and first age of active use through the KCT corpus (Lavalley *et al.* 2015).

Human language processing measures are based on research in cognitive science and information theory. They evaluate complexity in terms of processing costs associated with linguistic material, for example in terms of surprisal or cognitive load. We measure cognitive load in terms of integration costs based on dependency lengths and Gibson's (2000) Dependency Locality Theory (DLT). For the latter, we follow Shain *et al.*'s (2016) dependency parse based operationalization including variants which feature their suggested weight modifications for verbs, coordination, and modifiers.

3.2. System description

We extract our complexity features based on a three-step procedure. First, each text is linguistically annotated by applying a series of NLP tools and consulting external linguistic resources. In particular, texts are tokenized, segmented into sentences using OpenNLP 1.6.0 (http://opennlp.apache.org). Then, we perform PoS tagging, lemmatization, morphological analysis, and dependency parsing using the Mate tools 3.6.0 (Bohnet & Nivre 2012). We perform compound analysis using the JWordSplitter 3.4.0. (http://github.com/danielnaber/jwordsplitter). Finally, we obtain constituency parses using the Stanford PCFG parser 3.6.0 (Rafferty & Manning 2008) and topological field parses using the Berkeley parser 1.7.0 (Petrov & Klein 2007). While for many of these tasks, other NLP tools could also be employed, the mentioned tools, as far as we are aware, perform close to the state of the art in terms of quality and speed so that an exploration of alternatives is beyond the scope of this paper. In general, we use the default models provided by the respective tools for German analyses, except for topological field parsing, for which we used the model trained by Ziai & Meurers (2018) because the default topological field model is not compatible with the latest version of the parser.

These linguistic annotations are used in the second step to identify all instances of linguistic constructions that are relevant for our complexity analysis. This step relies on the identification of different linguistic units, some of which have different justifiable operationalizations, such as t-units or lexical words. To allow for comparability across complexity studies, it is crucial to make the underlying definitions of these units explicit (Bulté & Housen 2014; Housen *et al.* 2012). An elaborate documentation of the units underlying our system may be found in Weiss (2017: 82f). In the final step of the analysis, ratios and features are calculated to approximate the complexity of each document by means of a feature vector.[2] These are exported into a CSV table including all documents, which may then be used for further statistical evaluation.

To the best of our knowledge, this is currently the most extensive feature set for German complexity assessment. We are in the final stages of making the system publicly accessible via the Common Text Analysis Platform (CTAP) by Chen & Meurers (2016), which originally only facilitated English complexity analyses.

4. Merlin data

We analyze the non-normalized German section of the Merlin corpus (Abel *et al.* 2014) to assess German L2 writing proficiency. It is comprised of 1,033 texts written by the same number of adult learners of German, which have been elicited in official standardized language certification tests for the five CEFR test levels A1 to C1. With this, it is not only to our knowledge the largest freely available German L2 corpus, it also features text from an extraordinarily broad variety of thoroughly and transparently established proficiency levels. The corpus consists of approximately 200 texts per test level, which were prompted by overall 15 different tasks (three tasks per level). All texts are rated based on the CEFR scale from levels A1 to C2 by human experts for various performance categories as well as a holistic overall proficiency rating (Abel *et al.* 2014). Since learners achieved not only proficiency scores at the level of the test they took, but also scores above or below, the uniform distribution of test levels in Merlin does not translate to a uniform distribution of proficiency scores. Due to the negligible number of C2 rated texts (4 in total), we combined C1 and C2 texts to a single C1/C2 level rating for the purposes of our statistical analyses.

2 All features and formulas available at http://www.sfs.uni-tuebingen.de/~zweiss/rsrc/feat.pdf

5. Classification study

5.1. Set-up

As classification algorithm for our assessment of overall L2 proficiency, we chose the Sequential Minimal Optimization (SMO) support vector classifier by Platt (1998) with a linear kernel. This state-of-the-art algorithm is known to be relatively robust for many, potentially correlated measures and thus particularly suited for our large sets of complexity measures. We applied the SMO algorithm to varying combinations of complexity features. First, we grouped features together that assess the same theoretical or psycholinguistic linguistic domain. This resulted in seven linguistically homogeneous, theory-based feature groups: clausal, phrasal, lexico-semantic, and morphological complexity, discourse, human language processing (HLP), and language use (LU). Second, we performed information gain ranking to identify data-driven the most 50, 100, 150, and 200 informative features across all seven linguistic domains. We then discarded all but the most successful classifier, *IG 150*, which uses the 150 most informative complexity measures. Finally, we trained a classifier using all features. All classifiers were trained and tested using 10-fold cross-validation. The information gain ranking, too, was performed with this method. For further comparison, we also obtained a majority baseline by automatically assigning the most frequent proficiency level (B2) to all texts. We used the WEKA machine learning toolkit (Hall *et al.* 2009) for all analyses.

5.2. Results and discussion

5.2.1. Overall classification performance

Table 1 shows the overall performance of the different proficiency classifiers. All of them clearly outperform the majority baseline (32.0%). The best performing classifier is *IG 150*. Compared to the model using all features, removing less relevant features seems to slightly increase classification performance. More importantly, however, *IG 150* clearly outperforms the linguistically homogeneous classifiers yielding accuracies between 53.7% (HLP) to 67.6% (lexico-semantic).

Model	Num. features	Accuracy
Majority baseline	1	32.0
All features	400	68.1
IG 150	150	70.0
Discourse	84	64.7
Clausal	110	63.8
Phrasal	41	62.1
Lexico-semantic	38	67.6
Morphological	39	59.7
HLP	32	53.7
Language use	54	59.3

Table 1. Classification performance of SMO models in 10-fold cross-validation

Table 2 shows the confusion matrix for *IG 150*. Columns represent the proficiency scores predicted by the model, rows the actual proficiency scores assigned to a text in the Merlin corpus. For each observed score the most often predicted score was marked with bold font.

Obs.\Pred.	A1	A2	B1	B2	C1/C2	∑ Obs.
A1	21	**35**	1	0	0	57
A2	13	**231**	62	0	0	306
B1	1	50	**218**	62	0	331
B2	0	3	37	**252**	1	293
C1/C2	0	0	1	**44**	1	46
∑ Pred.	35	319	319	358	2	1,033

Table 2. Confusion matrix for IG 150

CEFR levels A2, B1, and B2 show favorable classification results: most predictions for texts from these levels are correct. For levels A1 and C1/C2, however, miss-classifications with their adjacent level are more common than correct classifications. This issue is particularly severe for level C1/C2. There are several potential explanations for this issue: partially it might be an artifact of the skewed distribution of Merlin overall CEFR scores and the resulting under-representation of these two levels: less than 10% of the corpus contains data with overall CEFR scores at levels A1 and C1/C2. For level A1, the classification might also suffer from the highly non-standard language of beginning learners, which impairs the automatic NLP annotations on which the

complexity features are based. Tono (2013) observes a risk-taking phase reaching into the intermediate level, where the number of errors increases together with complexity. While this could also cause problems for the NLP analysis, there is no indication for this in our results given the high classification accuracy for intermediate learners. For level C1/C2 another plausible explanation would be that the differences between B2 and C1 learners relate more to phraseological and stylistic writing aspects (Paquot 2017; Biber *et al.* 2011), which are not sufficiently captured in the current set of complexity features.

Rank	Avg. merit	Feature	Group
1	0.889 ± 0.010	Number of tokens	Descriptive
2	0.827 ± 0.019	Corrected type token ratio	Lexical
7	0.466 ± 0.009	Longest word in syllables	Lexical
8	0.432 ± 0.015	Sum of longest dependencies per sentence	HLP
13	0.391 ± 0.011	Dep. clauses with conjunction per dep. clause	Clausal
14	0.391 ± 0.006	Coverage of NP modifier types	Phrasal
16	0.387 ± 0.009	Dependent clauses per sentence	Clausal
22	0.372 ± 0.009	P(not-not) per transition	Cohesion
25	0.369 ± 0.012	Verbs per sentence	Phrasal
27	0.359 ± 0.025	VP modifiers per VP	Phrasal
29	0.358 ± 0.015	Words per t-unit	Clausal
31	0.355 ± 0.007	Sum non-terminal nodes per word	Clausal
35	0.354 ± 0.013	Standard deviation of verb cluster sizes	Phrasal
36	0.350 ± 0.007	P(not-object) per transition	Cohesion
37	0.350 ± 0.005	To-infinitives per sentence	Phrasal
39	0.346 ± 0.009	Total integration cost at finite verb per finite verb (with additional verb weight)	HLP
43	0.344 ± 0.006	HDD	Lexical
44	0.341 ± 0.011	Syllables per word	Lexical
50	0.326 ± 0.008	Temporal (Eisenberg) connectives per sentence	Cohesion
52	0.324 ± 0.013	Coverage of verb cluster sizes	Phrasal

Table 3. Top 20 complexity measures based on 10-fold cross-validated information gain with Pearson correlation less extreme than r ± 0.7

To examine closer the best performing classifier, *IG 150*, Table 3 shows the 20 most informative features included in the model. To allow for a broader view on the features represented in the model, we excluded measures which showed an extremely high Pearson correlation with higher ranking measures, i.e. that were more extremely correlated than ±0.7. The table shows the original total rank of each feature and their average merit in the 10-fold cross validation to allow for a more informed comparison of their overall informativeness. It also includes a reference to the feature group each feature is attributed to.

The results confirm that our data-driven feature selection approach in fact yields a highly diverse collection: the features include measures from nearly all feature groups and include operationalizations of the elaborateness and variation of these domains. The elaborateness of clausal subordination, the elaborateness and variation of nominal and verbal modification, lexical diversity and sophistication, transitions of grammatical roles and temporal connectives, and dependency-length based cognitive integration costs are particularly informative. Features of language use and morphological complexity are not represented in Table 3. However, they are repeatedly represented among the most informative 50 not extremely correlated features. Furthermore, morphological complexity features are represented among the higher-ranking measures, but highly correlated with word length and corrected type token ratio thus not eligible for Table 3. This holds in particular for derivational measures indicating nominalizations. The informativeness of language use measures is partially impaired by the type of data that is being analyzed: since we do not investigate the normalized texts and misspelled words will not be found in any of our word frequency data bases.

5.2.2. Classification performance by proficiency level

In the last step of our analysis, we investigated the relevance of certain linguistic domains and feature combinations for the identification of individual proficiency level changes with increasing proficiency. For this, we compared the performance of all classifiers for each individual proficiency level in terms of precision, recall, and f1 score as displayed in Table 4.

	A1			A2			B1			B2			C1/C2		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Maj	0.0	0.0	0.0	0.0	0.0	0.0	32.0	100	48.5	0.0	0.0	0.0	0.0	0.0	0.0
All	45.7	36.8	40.8	68.9	71.6	70.2	66.3	65.3	65.8	72.1	76.8	74.4	38.7	26.1	31.2
150	60.0	36.8	45.7	72.4	75.5	73.9	68.3	65.9	67.1	70.4	86.0	77.4	50.0	2.2	4.2
LU	100	1.8	3.4	63.5	63.7	63.6	52.7	56.2	54.4	62.1	78.8	69.5	0.0	0.0	0.0
HLP	0.0	0.0	0.0	58.5	61.8	60.1	47.2	51.1	49.1	56.0	67.2	61.1	0.0	0.0	0.0
Disc	66.7	3.5	6.7	66.1	73.2	69.5	63.0	54.1	58.2	64.6	89.8	75.1	0.0	0.0	0.0
Cla	50.0	7.0	12.3	63.5	75.2	68.9	59.7	57.4	58.6	68.1	80.2	73.7	0.0	0.0	0.0
Phr	33.3	1.8	3.3	62.7	74.8	68.3	56.9	55.9	56.4	66.8	77.5	71.7	0.0	0.0	0.0
Lex	100	15.8	27.3	68.6	78.4	73.2	65.3	59.2	62.1	67.6	86.3	75.9	0.0	0.0	0.0
Mor	0.0	0.0	0.0	58.7	68.3	63.1	53.6	54.5	54.0	67.1	77.8	72.0	0.0	0.0	0.0

Table 4. Precision, recall, and f1 score for homogeneous SMO models

Since the majority baseline labels all texts as level B2, its performance scores are zero for all but this level. More interestingly, none of the linguistically homogeneous models correctly identifies a single instance of C1/C2 level writings. Only the linguistically diverse models correctly classify some of these texts. The classifier using all measures even outperforms *IG 150* for this level, but worse on all others. This might indicate that there are some relevant features for the distinction of B2 and C1/C2 level writing, which were not elicited by the data-driven feature selection due to the under-representation of C1/C2 level texts. Potentially for similar reasons, for level A1 only very few texts are correctly classified, which results in extremely high precision scores with very low recall. The highest f1 (= 27.3%) score is achieved by the lexical complexity model. In contrast, HLP and morphological measures do not classify any text as A1.

For levels A2, B1, and B2 classification performance is better, although levels A2 and B2 exhibit systematically higher recall than precision values. An investigation of the misclassifications confirmed that these are predominantly due to the incorrect labeling of A1 and C1/C2 texts. Across these three levels, the lexical complexity model again performs best in terms of f1 score. Furthermore, clausal and phrasal complexities as well as discourse are considerably more successful at identifying texts at these levels than the other feature groups. For level B2 texts, the discourse model performs comparable to the lexical model. In contrast, HLP is the least informative feature group across all levels. This might be due to the limited diversity within this feature group, which predominantly consists of differently weighted instances of DLT integration cost measures. Language use also performs relatively poorly across

these levels. Morphological complexity shows little discriminatory power for most proficiency levels. Interestingly, however, it is as successful as phrasal and clausal complexities for the B2 level. The high relevance of discourse and morphological measures for B2 level texts is remarkable. Since the distinction between B2 and C1/C2 fails for the homogeneous models, this shows that the morphological and discourse models are able to learn a systematic distinction between B2 and the lower levels, while the discourse, clausal, phrasal, and lexical models identify systematic differences across levels A2 to B2.

6. Outlook

Our study clearly demonstrates the benefits of modeling L2 proficiency using broad, linguistically diverse feature selections and yields interesting insights regarding performance differences of linguistically homogeneous classifiers at individual proficiency levels. Yet, it also raised some issues that could only briefly addressed in our current study and that need further investigation. In particular, correctly identifying level A1 and level C1/C2 texts remains a challenge to all presented classifiers. This issue is partially due to the lack of data support for these proficiency levels in our data set. In a follow-up study, the extent to which our results are influenced by this artifact of the data distribution in Merlin needs to be investigated by additional analyses on more balanced data subsets.

Furthermore, there are two additional potential influences on our results that might also contribute to this issue and which we are currently addressing in ongoing studies. On the one hand, our models do not account for nonlinear development of individual measures. This work presented a view on various feature groups to illustrate the relevance of broad language modeling for proficiency assessment and to identify differences in the overall impact of linguistic domains on proficiency assessment. Building on this, we have moved on to the analysis of individual, potentially nonlinear measures by training Generative Additive Models (GAMs) on the Merlin data. In Weiss (2017), we present first results of this approach, where we closely model 13 complexity measures from all our feature groups including nonlinear developments. The results show an improvement in the classification of levels A1 and C1/C2 compared to the models presented here.

On the other hand, the Merlin data entails a particularly broad task background with three elicitation tasks per test level. These may cause task effects in the linguistic properties of the learner texts, in particular with regard to their

complexity, as earlier research on task effects and language performance has shown (Alexopoulou *et al.* 2017; Yoon & Polio 2016; Tracy-Ventura & Myles 2015). Thus, we broaden our investigation of the applicability of diverse complexity features to their sensitivity to task effects in learner corpora, thus shifting the focus of our analysis from learners to tasks. We have analyzed Merlin's elicitation tasks for various functional and cognitive task factors and performed first analyses on the effect these factors have on individual complexity measures as well as feature groups (Weiss 2017). Our preliminary results show that some complexity measures and groups seem to be sensitive to task effects to varying degrees: morphological complexity, for example, is particularly susceptible to task effects, while human language processing features seem to be remarkably robust (see Weiss 2017 for details). Both of these analysis strands have already yielded promising results and are currently pursued further.

7. Conclusion

We investigated to which extent broad linguistic modeling is beneficial for German L2 proficiency assessment. For this, we automatically extracted 400 measures of linguistic complexity from various linguistic domains with an elaborate NLP pipeline. We focused on comparing feature groups. On the one hand, we combined features from various linguistic domains in a data-driven approach. On the other, we grouped features together from the same linguistic domain. We compared them in terms of their ability to successfully distinguish between five holistic CEFR proficiency scores assigned to German L2 writings (A1 to C1/C2) when employed in SMO classifiers. Our results show that a broad selection of features that integrates aspects of language as a system, language use, and human sentence processing costs, results in higher classification performance on language learner data. In particular, lexical variation, sentential elaboration, phrasal elaboration and variation, and discourse elaboration are highly beneficial, as an analysis of the overall most informative measures in terms of information gain showed.

In a second step, we investigated to which extent the relevance of certain linguistic domains for the identification of individual proficiency levels changes with increasing proficiency. For this, we compared the performance of the classifiers assessing certain linguistic domains for identifying each individual proficiency level. This showed that lexical, clausal, and phrasal complexity are informative for the identification of several proficiency levels. In contrast, morphological and discourse measures are mostly relevant for distinguishing

B2 from lower proficiency levels. Human language processing and language use features are less successful, although we found individual measures from both groups to be highly informative and included in the classifier using features from various domains. In this analysis, too, the combination of features outperformed all linguistically homogeneous models across individual proficiency levels. Overall, our results show that broad linguistic modelling is beneficial and feasible for German L2 proficiency assessment, even when applied to non-normalized data.

References

Abel, A., Wisniewski, K., Nicolas, L., Boyd, A., Hana, J. & Meurers, D. (2014). A trilingual learner corpus illustrating european reference levels. *Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne* 2(1), 111-126.

Alexopoulou, T., Michel, M., Murakami, A. & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1), 180-208.

Barzilay, R. & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics* 34, 1-34.

Biber, D., Gray, B. & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1), 5-35.

Bohnet, B. & Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455-1465.

Bulté, B. & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26, 42-65.

Chen, X. & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 113-119.

Crossley, S., Kyle, K., Allen, L., Guo, L. & McNamara, D. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment* 7(1).

Crossley, S. & McNamara, D. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2), 115-135.

Crossley, S. & McNamara, D. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26, 66-79.

Ellis, R. & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

François, T. & Fairon, C. (2012). An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466-477.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita & W. O'Neil (eds) *Image, Language, Brain*. Cambridge: MIT Press, 95-12.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* 11(1), 10-18.

Hancke, J., Vajjala, S. & Meurers, D. (2012). Readability classification for German using lexical, syntactic and morphological features. *Proceedings of COLING 2012: Technical Papers*, 1063-1080.

Hancke, J. & Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. *Book of Abstracts of the Learner Corpus Research Conference 2013, Universitetet i Bergen, Norway*, 54-56.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A. & Kliegl, R. (2011). dlexDB - Eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau* 62(1), 10-20.

Henrich, V. & Hinrichs, E. (2010). GernEdiT - The GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 2228-2235.

Housen, A., Vedder, I. & Kuiken, F. (eds). (2012). *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*. Amsterdam & Philadelphia: John Benjamins.

Lavalley, R., Berkling, K. & Stüker, S. (2015). Preparing children's writing database for automated processing. In *Language Teaching, Learning and Technology (LTLT-2015), Leipzig, 4 September, 2015*, 9-15.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474-496.

McNamara, D., Crossley, S. & McCarthy, P. (2009). Linguistic features of writing quality. *Written Communication* 27(1), 58-86.

Norris, J. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555-578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492-518.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (eds) *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: de Gruyter, 127-155.

Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 1-25.

Petrov, S. & Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL-HLT 2007, Rochester, 22-27 April, 2007*, 404-411.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Reports MSR-TR-98-14*.

Rafferty, A. & Manning, C. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *ACL Workshop on Parsing German (PaGe-08), Columbus, 20 June, 2008*, 47-54.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E. & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 49-58

Taguchi, N., Crawford, W. & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly* 42(2), 420-430.

Tono, Y. (2013). Criterial feature extraction using parallel learner corpora and machine learning. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds)

Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam & Philadelphia: John Benjamins, 169-204.

Tracy-Ventura, N. & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research* 1(1), 58-95.

vor der Brück, T., Hartrumpf, S. & Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica* 32, 429-435.

Weiss, Z. (2017). *Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects.* Master thesis, University of Tübingen. http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/ weiss2017-distr.pdf (last accessed on 26 September, 2018).

Yoon, H.-J. & Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly* 51(2), 275-301.

Ziai, R. & Meurers, D. (2018). Automatic focus annotation: Bringing formal pragmatics alive in analyzing the information structure of authentic data. In *Proceedings of NAACL-HLT 2018*, 117-128.

Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School

Zarah Weiss and **Detmar Meurers**
University of Tübingen
Department for General and Computational Linguistics
{zweiss,dm}@sfs.uni-tuebingen.de

Abstract

We track the development of writing *complexity* and *accuracy* in German students' early academic language development from first to eighth grade. Combining an empirically broad approach to linguistic complexity with the high-quality error annotation included in the Karlsruhe Children's Text corpus (Lavalley et al., 2015) used, we construct models of German academic language development that successfully identify the student's grade level. We show that classifiers for the early years rely more on accuracy development, whereas development in secondary school is better characterized by increasingly complex language in all domains: linguistic system, language use, and human sentence processing characteristics. We demonstrate the generalizability and robustness of models using such a broad complexity feature set across writing topics.

1 Introduction

We model the development of linguistic complexity and accuracy in German early academic language and writing acquisition from first to eighth grade. Complexity and Accuracy are well-established notions from Second Language Acquisition (SLA) research. Together with Fluency, they form the CAF triad that has successfully be used to characterize second language development (Housen et al., 2012). Accuracy here is defined as a native-like production error rate (Wolfe-Quintero et al., 1998) and Complexity as the elaborateness and variation of the language which may be assessed across various linguistic domains (Ellis and Barkhuizen, 2005).

While there has been substantial research on the link between linguistic complexity analysis and second language proficiency and writing development for English (cf., e.g., Bulté and Housen, 2014; Kyle, 2016), much less is known about academic language development for other languages, such as the morphologically richer German. In this article, we target this gap with three contributions. We build classification models for early academic language development in German from first to eighth grade, based on a uniquely broad set of linguistically informed measures of complexity and accuracy. Our results indicate that two phases of academic language development can be distinguished: Initial academic language and writing acquisition focusing on the writing process itself, best characterized in terms of accuracy development, with little development in terms of complexity. A second stage is characterized by the increasing linguistic complexity, in particular in the domains of lexis and syntactic complexity at the phrasal level. We demonstrate the robustness and generalizability of the models informed by the broad range of linguistic characteristics – a major concern not only for obtaining practically relevant approaches for real-life use, but also for characterizing machine learning going beyond focused task to approaches capable of capturing general language characteristics.

The article is structured as follows: We first give a brief overview of research on writing development in terms of complexity and accuracy. We then present the *Karlsruhe Children's Text* corpus used as empirical basis of our work. In Section 4, we spell out our approach to assessing writing in terms of complexity and accuracy, before sections 5, 6, and 7 report on three studies designed to address the research issues introduced above.

2 Related Work

The main strand of research analyzing the complexity and accuracy constructs targets the assessment of second language development. Linguistic complexity measures have been successfully used to model the language acquisition of English

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 380–393
Florence, Italy, August 2, 2019.

as a Second Language (ESL) learners (Bulté and Housen, 2014; Crossley and McNamara, 2014). Work on first language writing development for English has also been conducted, but it is less common (Crossley et al., 2011) The same holds for the development of accuracy (Larsen-Freeman, 2006; Yoon and Polio, 2016). Most studies focus on adult ESL learners' development during periods of instruction. Vercellotti (2015) finds an increase in syntactic and lexical complexity, overall accuracy, and fluency in adult ESL speech over the course of several months. Crossley and McNamara (2014) find that advanced adult ESL learners phrasal and clausal complexity significantly increases over the course of one semester of writing instruction in particular with regard to nominal modification and number of clauses. These findings are corroborated by Bulté and Housen (2014). For uninstructed settings, however, this does not hold. Knoch et al. (2014, 2015) study university students' ESL development over 12 months and three years without instruction in an immersion context and found that only fluency but not grammatical and lexical complexity developed.

Research on languages other than English is starting to appear (Hancke et al., 2012; Velleman and van der Geest, 2014; Pilán and Volodina, 2016; Reynolds, 2016). As for English, research on German writing development has predominantly focused on German as a Second Language (GSL) in instructed settings (Byrnes, 2009; Byrnes et al., 2010; Vyatkina, 2012). Their findings suggest that as for ESL learners' writing, clausal complexity progressively increases. For lexical complexity results have been more mixed depending on the proficiency of GSL learners' proficiency level. The development of writing accuracy has also been assessed in some corpus studies using automated or manual error annotation (Lavalley et al., 2015; Göpferich and Neumann, 2016). In Weiss et al. (2019) we analyze the impact of linguistic complexity and accuracy on teacher grading behavior.

One challenge for the assessment of language performance in terms of complexity that is starting to receive attention is the influence of the task. Alexopoulou et al. (2017) demonstrate task effects, specifically task complexity and task type, on the complexity of English as a Second Language writers in the EF-Cambridge Open Language Database (EFCAMDAT) and show mixed results for accuracy. This is in line with findings by Yoon and Polio (2016), who investigate the effect of genre differences on CAF constructs. Yoon (2017) focuses on the effect of topic on the syntactic, lexical, and morphological complexity of ESL learners' writings and shows a significant influence on the complexity of writings of the same learners, similar to findings in Yang et al. (2015). Such task effects have mostly been discussed from a theoretical perspective, considering their implications for the development of CAF constructs and the two main task frameworks (Robinson, 2001; Skehan, 1996). From a more practical perspective, task, genre, and topic effects have been recognized as an important issue for machine learning for readability assessment or Automatic Essay Scoring (AES). For the real-world applicability of such approaches it is crucial for them to account for differences due to genre or topic. In their readability assessment system *READ-IT* for Italian, Dell'Orletta et al. (2014) use this issue to motivate favoring a ranking-based over a classification-based approach. A recent AES approach discussing the issue is the placement system for ESL by Yannakoudakis et al. (2018).

3 Data

Our studies are based on the *Karlsruhe Children's Text* (KCT) corpus by Lavalley et al. (2015).[1] It is a cross-sectional collection of 1,701 German texts produced by students in German elementary and secondary school students from first to eighth grade. The secondary school students in the corpus attended one of two German school tracks, either a basic school track (*Hauptschule*) or an intermediate school track (*Realschule*). The texts were written on a topic chosen by the students from a set of age-appropriate options: Elementary school students were asked to continue one of two stories, one about children playing in a park, and the other about a wolf who learns how to read. Secondary school students wrote about a hypothetical day spent with their idol or their life in 20 years. All student texts in the corpus are made available in the original, including all student errors, and a normalized version, where errors and misspellings were corrected. The data is enriched with error annotations covering word splitting, incorrect word choices and repetitions, grammar, and legibility.

For our studies analyzing writing development

[1] https://catalog.ldc.upenn.edu/LDC2015T22

in terms of development across the grade levels, we made use of the normalized texts and the error annotation. Some grade levels in the corpus include only few texts, such as the 42 cases of first grade writings compared to the other grade levels with 189 to 283 writings. We thus grouped adjacent grade levels, i.e., grades 1 and 2 together, grades 3 and 4, etc., to obtain a data set with a substantial number of instances for each class.

4 Assessment of Writing Performance

To assess writing performance in terms of complexity and accuracy, we operationalized these SLA concepts in terms of several features which we automatically computed or derived from the error annotation of the KCT corpus.

4.1 Complexity

The analysis of complexity is based on our implementation of a broad range of complexity features for German (Weiss, 2017; Weiss and Meurers, 2018, in press). The features cover clausal and phrasal syntactic complexity, lexical complexity, discourse complexity, and morphological complexity. Complementing the measures of complexity of the linguistic system, we also compute two cognitively-motivated features: a characterization of language use based on word frequencies, and measures of human language processing (HLP). Table 1 summarizes the features designed to capture the elaborateness and variability in the respective domain, with more details provided in Weiss (2017) and Weiss and Meurers (in press). Overall, the studies in the current paper make use of a comprehensive set of 308 complexity features for the assessment of academic language development.[2]

4.2 Accuracy

The second dimension of language performance that we are interested in is writing accuracy. In SLA research accuracy has predominantly been assessed in terms of types of error rates or error-free T-units (Wolfe-Quintero et al., 1998; Verspoor et al., 2012). We exploited the KCT corpus' elaborate error annotation to extract a broad range of accuracy measures. Annotations on the level of individual letters and words mark (ill)legibility, word splitting errors, repetition errors, foreign words, and grammatical errors. Annotations at the sentence level mark content deletions, insertions, and incorrect word choices. In addition, we developed an approach to automatically derive additional error types by comparing the original student writings with their normalized sentence-aligned target hypotheses. This procedure allowed us to extract counts for punctuation errors, incorrect quotation marks, spelling mistakes, and word capitalization errors. The last item is a particular challenge of German orthography, given that capitalization in German is governed by a complex set of rules and conventions relating to syntactic structure.[3]

Overall, we extracted 20 accuracy counts which we aggregated and normalized by the total number or errors or the total number of words in the text as counted by the complexity analysis described in the previous subsection. The feature set measuring writing accuracy and an example feature is included as the last row in Table 1.[4]

5 Study 1: Predicting Grade-Levels across School Types

5.1 Set up

We extracted the text data from the KCT corpus, removing all texts containing less than ten words and excluding texts written by children younger than seven years and older than 15 years. This resulted in a corpus of N=1,633 texts, for which we computed the features of linguistic complexity and error rates. Table 2 shows the distribution of texts across grade levels and school tracks.

From the analyzed data set, we eliminated all complexity and error rate features that did not exhibit enough variability to be of interest for the analysis. Specifically, we excluded all features whose most common value occurred more than 90% of the time. For the remaining 262 features, we computed their z-score, centered around zero.

On this data, we performed ten iterations of 10-fold cross-validation (CV) generating different splits each time, i.e., 100 training and testing runs in total, using an SMO classifier with a linear kernel (Platt, 1998). This outperformed models using random forests or linear regression. Similarly, introducing non-linearity did not improve the clas-

[2] We are making the complexity code available as part of a multilingual version of CTAP: https://github.com/zweiss/multilingual-ctap-feature

[3] The Python script used to identify accuracy features in the KCT annotation is available at https://github.com/zweiss/KCTErrorExtractor

[4] Here and in the following, we will refer to this feature set as the *error rate* measures to avoid confusion with the term accuracy used as a classification performance metric.

Feature Set	Size	Description
Lexical complexity	31	measures vocabulary range (lexical density and variation) and sophistication, measures of lexical relatedness; e.g., type token ratio
Discourse complexity	64	measures the use of cohesive devices such as connectives; e.g., connectives per sentence
Phrasal complexity	47	measures of phrase modification; e.g., NP modifiers per NP
Clausal complexity	27	measures of subordination or clause constituents; e.g., subordinate clauses per sentence
Morphological complexity	41	measures inflection, derivation, and composition; e.g., average compound depth per compound noun
Language Use	33	measures word frequencies based on frequency data bases; e.g., mean word frequency in Subtlex-DE (Brysbaert et al., 2011)
Human Language Processing	24	measures of cognitive load during human sentence processing, mostly based on Dependency Locality Theory (Gibson, 2000) e.g., average total integration cost at the finite verb
Error Rate	41	measures ratios of error types per error or word; e.g., spelling mistakes per word

Table 1: Overview over the feature sets used to capture linguistic complexity and accuracy

	1/2	3/4	5/6	7/8	all
Elementary	203	524	0	0	727
Realschule	0	0	297	236	533
Hauptschule	0	0	165	208	373
all	203	524	462	444	1633

Table 2: Text distribution across grades & school tracks

sification. For each feature set introduced in Section 4, we trained a separate classifier to support a comparison of the different complexity and error feature sets. In addition, we built one classifier based on the combination of all complexity feature sets and one combining all feature sets including error rate. Finally, we built a classifier also including the meta information about the school track and topic chosen, to investigate their influence on the complexity features and the comparability of grade-levels across school types.

As reference for evaluating classifier performance, we use a majority baseline assigning always the most common grade level, and a second baseline inspired by traditional readability formulas, for which we trained a classifier using text length and average word length features.

5.2 Results & Discussion

Table 3 shows the performance of the classifiers in terms of mean accuracy and standard deviation across iterations and folds, and the feature set size. The majority baseline and the traditional readability feature baseline displayed above the dashed line are both around 32%. All linguistically informed classifiers clearly outperform these two baselines. The best performing model with an accuracy of 72.68% combines linguistic complexity features and error rate with information on topic and school track.[5] Adding this meta-information, which in most real-life application contexts is readily available, accounts for an 1.72% increase in accuracy. But also without this meta-information, the combination of linguistic complexity features and error rate is highly successful with an accuracy of 70.96%.

Let us take a look at the individual contributions of the different feature sets. The overall linguistic complexity classifier clearly outperforms the one informed by the error rate features. This comparison may be biased towards the linguistic complexity classifier because it is informed by six times more features. However, the impression that complexity features are more indicative for writing development as a function of grade level is supported by the classifiers based on individual domains of linguistic complexity, which are more comparable in size to the error rate based classifier. The lexical complexity, discourse complexity, and phrasal complexity classifiers all clearly outperform the classifier informed by error rate with accuracies between 60.10% and 61.29% compared to 54.47%. The same holds for morphological

[5] The confusion matrix for all ten iterations of the 10-CV may be found in Table 10 in the Appendix.

	Size	μ-Acc.	SD-Acc.
Majority baseline	1	32.08	0.14
Traditional baseline	2	32.56	0.80
All Features + Meta	264	**72.68**	1.94
All Features	262	**70.96**	2.01
Complexity	225	**68.35**	2.25
Error Rate	37	54.47	2.11
Lexical	31	60.10	1.69
Discourse	48	60.10	1.66
Phrasal	41	61.29	1.73
Clausal	26	52.95	1.56
Morphological	27	56.45	1.47
Language Use	30	45.45	1.28
Human processing	20	42.18	1.55

Table 3: Grade-level classification of elementary & secondary school texts, ten iterations of 10-fold CV, distinguishing levels 1st/2nd, 3rd/4th, 5th/6th, 7th/8th

complexity (56.45%), although the difference is less pronounced. However, not all dimensions of linguistic complexity outperform error rate. This holds only for features measuring the linguistic system. While psycho-linguistic measures of language use and human language processing clearly outperform the baselines, they are performing significantly worse than the error rate features. Language experience and cognitive measures of the complexity in processing language does not seem to be the factor limiting academic writing performance, which is intuitively plausible considering that, especially in the early school years, the language experience and language processing will be mostly shaped by spoken language interaction.

6 Study 2: Writing Development in Elementary vs. Secondary School

6.1 Set-Up

Having established that linguistic complexity and error rate successfully predict writing performance across academic writing development, let us compare the development in early writing with that in secondary school. For this, we split the KCT data into two subsets: one containing only elementary school writing ($N = 727$), the other the secondary school writing from the different school tracks ($N = 906$). We applied the same pre-processing steps described in Section 5.1 including feature reduction and scaling of all predictor variables, obtaining 256 features for the elementary school and 255 for the secondary school data set (with numbers differing slightly since the feature reduction is performed separately on each data set).

We then followed a two-fold approach: First, we again tested and trained the same SMO classifiers as in Study 1 with linear kernels and 10 iterations of 10-fold CV (Section 6.2). Although the classifiers were informed by the same feature sets, due to the reduction of the sample size some sets were reduced more in the aforementioned preprocessing step which may result in slightly deviating feature set sizes across tables. For the elementary school data set, only topic was added as meta information, because there are no different elementary school tracks in Germany.

Then, for both data sets we selected the most informative features of each feature set in order to zoom in on how they differ across grade-levels (Section 6.3). This more fine grained analysis allows us to complement the broader perspective gained form the classification experiments with a more concrete sense of which features matter and how they change. For the selection, we ranked all features by their information gain for the distinction of grade-levels in the respective data set and selected the most informative feature of each feature set resulting in overall 16 features chosen for closer inspection. We then conducted two-tailed t-tests to test for significant differences across grade-levels in both data sets. To avoid redundancy in our comparison, if the most informative feature for a given feature set in both data subsets assessed the same concept, we chose the next-most informative feature.[6]

6.2 Results & Discussion

Table 4 shows the classifiers performance on the elementary school data subset.

Unlike in the previous study, the majority baseline for this binary classification task is relatively high with 71.72% given that there is less data for the first and second grade. As in the first study, the second baseline using the traditional readability formula features text length and average word length performs only at the level of the majority baseline. The classifier combining evi-

[6] For example, the most informative feature of lexical complexity is in both subsets a measure of lexical diversity (Yule's k and root type-token ratio). Due to its higher ranking (overall most informative for secondary school) and its reduced sensitivity to text length, we chose to keep Yule's k and included the second most informative lexical complexity feature for elementary school: corrected verb variation (measuring lexical variation).

	Size	μ-Acc.	SD-Acc.
Majority baseline	1	71.72	0.35
Traditional baseline	2	71.72	0.35
All Features + Meta	256	**82.81**	2.11
All Features	255	**82.60**	1.97
Complexity	218	77.93	2.42
Error Rate	37	**81.56**	1.27
Lexical	31	77.32	1.92
Discourse	46	75.18	1.71
Phrasal	39	76.77	2.18
Clausal	26	72.44	0.49
Morphological	27	71.72	0.35
Language Use	30	71.72	0.35
Human processing	19	71.72	0.35

Table 4: Grade-level classification of elementary school texts, ten iterations of 10-fold CV, distinguishing levels *1st/2nd* and *3rd/4th*

dence from linguistic complexity features and error rate clearly outperforms the baselines with an accuracy of 82.60%.[7] Adding meta-information, which here means adding the writing topic, does not make a significant contribution.

Looking at the classifiers for the subsets of features, we see that error rate features make a significant contribution. While the difference in performance still is significant,[8] the classifier informed only by error rate features with an accuracy of 81.56% performs close to the combined model with an accuracy of 82.60%. The classifier using only complexity features performs worse, with an accuracy of 77.93%, even though this classifier is informed by considerably more features. When looking at the individual domains of linguistic complexity, again lexical complexity, discourse complexity, and phrasal complexity are the most informative features, but they perform significantly lower than the error rate features. The other domains of linguistic complexity seem to be uninformative for the grade level distinction in elementary school student writings – clausal and morphological complexity, language use, and human language processing all perform at baseline level.

Our findings show that early writing and academic language development predominantly focuses on establishing writing correctness rather than language complexification. However, in certain domains writing performance also advances in terms of complexity, namely the lexicon, discourse, and phrase complexity. Systematic improvements in the domains of clausal and morphological complexity or language use and human language processing, however, do not take place.

Turning to the secondary school data set, Table 5 shows the classification results for that subset.

	Size	μ-Acc.	SD-Acc.
Majority baseline	1	51.15	0.27
Traditional baseline	2	51.56	1.75
All Features + Meta	258	**65.66**	2.13
All Features	255	**63.71**	1.82
Complexity	220	**64.16**	1.63
Error Rate	35	54.34	2.48
Lexical	30	62.74	1.58
Discourse	45	57.13	1.75
Phrasal	41	57.64	2.10
Clausal	25	58.70	2.37
Morphological	27	54.31	2.39
Language Use	30	55.73	2.34
Human processing	18	52.67	1.90

Table 5: Grade-level classification on secondary school texts, ten iterations of 10-fold CV, distinguishing levels: *5th/6th* and *7th/8th*

The data set is more balanced across grouped grade levels, with a majority baseline of 51.15%. Traditional readability features again perform at the same level as the majority baseline. The best performing classifier again combines the features encoding linguistic complexity and error rate with information on topic and school track. It reaches an accuracy of 65.66%, performing nearly 2% better than the model without the meta-information.[9] Different from the elementary school data classifier, we here also distinguish the two secondary school tracks, which apparently differ in the complexity of the texts written in a given grade level.

A comparison of the classifiers based on error rate features versus the complexity features shows that for secondary school grade levels linguistic complexity is more indicative for differentiating grade levels. The classifiers differ in terms of their accuracy by nearly 10%. When comparing the performance of error rate features with the individual domains of linguistic complexity, we see that this difference cannot merely be explained by

[7] The confusion matrix for all ten iterations of the 10-CV may be found in Table 11 in the Appendix.

[8] One-sided t-test: t = −4.3978, df = 169.34, p = 9.63e-06

[9] The confusion matrix for all ten iterations of the 10-CV may be found in Table 12 in the Appendix.

the difference in feature set size. Lexical complexity, in particular, but also discourse complexity, phrasal complexity, and clausal complexity significantly outperform error rate features. This clear development of clausal complexity in secondary school writing is another difference to the development of writing of elementary school students. Language use and morphological complexity also show more development and significantly outperform the baselines. Human language processing features do not show a significant development.

Summarizing the findings from Table 4 and Table 5, we saw that the early writing and academic language development seemed to predominantly focus on establishing writing correctness rather than complexification. However, despite this focus on correctness, writing performance exhibits also in early stages of writing acquisition advances in terms of linguistic complexity in the domains of lexicon, discourse, and phrasal complexity. Systematic improvements in the other domains of linguistic complexity only take place at later stages of writing development. The beginning of this trend may be seen in the evidence from secondary school writings, for which clausal complexity and to some extent also morphological complexity and language use become increasingly relevant. Lexical complexity, phrasal complexity, and discourse complexity develop throughout all stages of writing acquisition.

6.3 Zooming in on Individual Features

Table 6 shows the most informative features from each feature set, their group means across grade-levels in elementary and secondary school, and the results of the t-tests.[10] In the first step (Section 6.2), we found that error rate as well as lexical, phrasal, and discourse complexity develop in both, elementary and secondary school writing. Zooming in on these domains, we see that some features systematically develop throughout grade-levels. Overall error rate and capitalization errors are highly informative in both data sets and decrease significantly across all grade-levels. Similarly, for lexical complexity, lexical diversity measured by Yule's k significantly decreases with progressing grade-levels (from 217 in grade-level 1/2 to 128 in grade-level 7/8). However, not in all cases the results are as clear. Lexical variation measured as corrected verb ratio significantly increases from grade-levels 1/2 to 3/4 and 5/6 to 7/8. Yet, the lexical variation of grade-level 7/9 writing is closer to that of grade-level 3/4 than 5/6, leaving unclear to which extent we see systematic development in this subdomain of lexical complexity.

For discourse complexity, the transition probability of dropping the subject in a following sentence, i.e., not repeating it as, e.g., the subject or object, significantly decreases with increasing grade-level in elementary school, i.e., the discourse becomes more coherent. The probability remains stable at a low level in secondary school. There, discourse complexity seems to develop rather in terms of use of connectives such as temporal connectives which significantly increase with progressing grade-level, while showing inconclusive results for elementary school. The two most informative features from the domain of phrasal complexity behave similarly: The coverage of noun phrase modifiers for elementary school which significantly increases from grades 1/2 to grades 3/4 from 0.31 to 0.42 but stagnates around 0.52 in secondary school. For secondary school, it is represented by the ratio of verb modifiers per verb, which significantly increases across all grade-levels from 0.29 to 0.65.

In contrast to phrasal complexity, clausal complexity represented by conjunction clauses per sentence and verbs per t-unit does not significantly change throughout elementary school. However, it significantly increases in secondary school from 0.13 conjunction clauses per sentence to 0.18 and from 1.69 verbs per t-unit to 1.8. This is in line with our previous observation that elementary school writing rather develops in terms of phrasal but not clausal complexity, while clausal complexity gains importance in secondary school.

The same holds for morphological complexity and language use, which we found to only play a role in the development of secondary school writing. Accordingly, we do not see a significant difference in either across elementary school grade-levels for the most informative features of these domains. For secondary school writing, however, the number of derived nouns per noun significantly increases, indicating a stronger nominal style in students writing and we see a significant increase in vocabulary overlap with dlexDB, which consists of frequencies from news

[10] The appendix contains the information gain ranking for the 16 most informative features for both data sets, see Tables 15 and 16 as well as boxplots visualizing of all features across grade-levels, see Figures 2 to 1.

Feature name	Set	Elementary school				Secondary school			
		1/2	3/4	t	p	5/6	7/8	t	p
Overall errors / W	Error Rate	0.68	0.37	11.53	.000	0.28	0.22	5.60	.000
Corrected verb variation	Lexical	1.62	2.13	-11.55	.000	1.88	2.01	-3.03	.003
P(Subject → Nothing)	Discourse	0.15	0.10	3.40	.001	0.05	0.06	-1.35	.177
Avg. NP modifier types	Phrasal	0.31	0.42	-8.93	.000	0.52	0.52	-0.21	.831
Conjunction clauses / S	Clausal	0.11	0.13	-0.96	.339	0.13	0.18	-3.47	.001
Finite verbs / verb	Morph.	0.82	0.81	1.63	.105	0.71	0.70	0.88	.381
Pct. LW in Subtlex	Language Use	0.04	0.05	-1.71	.089	.085	.077	1.82	.069
DLT-IC (M) / finite verb	Human Processing	1.09	1.11	-1.96	.051	1.22	1.25	-1.65	.099
Capitalization errors / W	Error Rate	0.15	0.07	9.87	.000	0.05	0.04	5.61	.000
Yule's K	Lexical	217.	153.	7.21	.000	152.	128.	5.60	.000
Temp. connectives / S	Discourse	0.73	0.63	1.85	.066	0.47	0.62	-4.10	.000
Verb modifiers / VP	Phrasal	0.29	0.49	-4.85	.000	0.55	0.65	-2.86	.004
Verbs / t-unit	Clausal	1.67	1.57	-0.97	.333	1.69	1.81	-3.18	.002
Derived nouns / noun	Morph.	0.02	0.02	-0.38	.708	0.04	0.05	-2.66	.008
Pct. LW in dlexDB	Language Use	0.62	0.60	1.60	.111	0.60	0.63	-3.27	.001
($\sum$ max. dep.) / S	Human Processing	5.12	5.60	-2.64	.009	6.30	6.97	-4.59	.000

Table 6: Across-grade level group means of the most informative features of each feature set for distinguishing grade-levels in elementary school (above dashed line) and secondary school (below dashed line).

texts. This might indicate that language use becomes more similar to news language in secondary school, as dlexDB is based on news paper data.

Interestingly, for human language processing, there seems to be a marginally significant increase in DLT processing costs at the finite verb (with decreased modifier weight as defined in Shain et al. 2016) and a significant increase in the mean maximal dependency length per sentence across all grade-levels in elementary and secondary school.

	1/2	3/4	5/6	7/8	all
Wolf	133	353	0	0	466
Park	90	171	0	0	261
Future	0	0	332	333	665
Idol	0	0	130	111	241
all	203	524	462	444	1,663

Table 7: Distribution of grade levels across topics

7 Study 3: Cross-Topic Testing of Academic Language Development Across Topics

7.1 Set Up

In our final study, we want to test whether the results we obtained generalize across topics. Elementary school and secondary school students were both allowed to freely choose from two different topics for their writing as spelled out in Section 3. We used the two data subsets from Study 2, but additionally split them by topics, obtaining four data sets: i) elementary school: *Wolf* topic, ii) elementary school *Park* topic, iii) secondary school: *Future* topic, and iv) secondary school *Idol* topic. Table 7 shows the distribution of texts across grade levels and topics.

We used the data sets of *Wolf* topic writings and *Future* topic writings as training data sets and tested the resulting model on *Park* topic and *Idol* topic texts, respectively. We chose this set-up since the two test data sets are too small to allow for training and testing with reversed data sets. We do not use cross-validation here, because we specifically want to study transfer across different topics rather than just different folds. In the new set-up, we cross-topic trained and tested the SMO classifiers based on the combination of complexity and error rate features and separately for the error rate and for the complexity features. We compared the results against the majority baseline and the traditional readability baseline containing measures of text and word length. For the secondary school data, we trained one model with and one without meta information on school tracks.

7.2 Results & Discussion

Table 8 shows the cross-topic classification performance on elementary school students' writings.

Feature Set	Train	Test	Acc.
Majority baseline	*n.a.*	Park	65.52
Traditional baseline	Wolf	Park	65.52
All Features	Wolf	Park	76.63
Complexity	Wolf	Park	68.58
Error Rate	Wolf	Park	81.61

Table 8: Cross-topic results for elementary school data

The majority baseline for elementary school writings' on the *Park* topic is more balanced than the one for the *Wolf* topic. For both topics, 3rd/4th grade was the most common grade-level. Training on *Wolf* texts and testing on *Park* texts with the SMO classifier yields an accuracy of 76.63%. While this does constitute a drop in accuracy as compared to Study 2, which may at least partially be explained by the reduced size of the training data set, the model clearly generalizes across topics. When taking a closer look at the difference between the purely error rate-based informed classifier and the complexity feature based classifier, we see that both generalize across topics. However, error rate clearly outperforms the complexity features and in fact hardly drops in performance when compared to the results obtained in Study 2.[11] The better performance of the classifier informed by error rate compared to both complexity-based classifiers indicates that error rate is more robust across topics than complexity. It also further corroborates the particular importance of writing correctness for early writing and academic language development.

Table 9 shows the results of the classifiers for the secondary school writing.

Feature Set	Train	Test	Acc.
Majority baseline	*n.a.*	Idol	50.01
Traditional baseline	Future	Idol	43.15
All Features + Meta	Future	Idol	62.66
All Features	Future	Idol	59.33
Complexity	Future	Idol	59.34
Error Rate	Future	Idol	55.19

Table 9: Cross-topic results for secondary school data

Unlike for the elementary school data, grade-levels are more or less balanced across topics for this data set, leading to a majority baseline around 50%. As before, we see that all SMO classifier generalize across topics when training on the larger data set (*Future*) and testing on the smaller one (*Idol*). In line with their relative importance for this school level established in the second study, the complexity features play more of a role and interestingly generalize well, while the error rate measures known to play less of a role at this level of development are also less robust.[12]

8 Conclusion and Outlook

We presented the first approach modeling the linguistic complexity and accuracy in German academic language development across grades one to eight in elementary and secondary school. Our models are informed by a conceptually broad feature set of linguistic complexity measures and accuracy features extracted from error annotations. The computational linguistic analysis made it possible to empirically identify a shift in the developmental focus from accuracy as the primary locus of development in elementary school to the increasing complexity of the linguistic system in secondary school. Our results also show where both domains advance in parallel, in particular in the lexical complexity domain, which plays an important role throughout. Despite the emerging focus on complexity throughout secondary school, accuracy also continues to play a role. Investigating the generalizability of our results and the approach to complexity and accuracy development, we demonstrated the cross-topic robustness of our classifiers. The use of cross-topic testing to establish the robustness of machine learning models thus supports the applicability of language development modeling in real life.

These first results provide insights into the complexity and accuracy development of academic writing across the first eight years in German. Yet, they are based on the quasi-longitudinal operationalization of writing development as a function of grade level. Tracking genuine longitudinal develop of individual students across extended periods of time is a natural next step, which will make it possible to study individual differences and learning trajectories rather than overall group characteristics. We plan to follow up on this in future work.

[11] The confusion matrix for all ten iterations of the 10-CV may be found in Table 13 in the Appendix.

[12] The confusion matrix for all ten iterations of the 10-CV may be found in Table 14 in the Appendix.

References

Theodora Alexopoulou, Marije Michel, Akira Murakami, and Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67:181–209.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58:412–424.

Bram Bulté and Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26(0):42 – 65. Comparing perspectives on L2 writing: Multiple analyses of a common corpus.

Heidi Byrnes. 2009. Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education*, 20(1):50 – 66.

Heidi Byrnes, Hiram H. Maxim, and John M. Norris. 2010. Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment. *The Modern Language Journal*, 94.

Scott A Crossley and Danielle S McNamara. 2014. Does writing development equal writing quality? a computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26:66–79.

Scott A. Crossley, Jennifer L. Weston, Susan T. McLain Sullivan, and Danielle S. McNamara. 2011. The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3):282–311.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of the International Journal of Applied Linguistics*, 165(2):163–193.

Rod Ellis and Gary Barkhuizen. 2005. *Analysing learner language*. Oxford University Press.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

Susanne Göpferich and Imke Neumann. 2016. Writing competence profiles as an assessment grid? – students' L1 and L2 writing competences and their development after one semester of instruction. In *Developing and Assessing Academic and Professional Writing Skills*, pages 103–140. Peter Lang, Bern, Switzerland.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Alexis Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Ute Knoch, Amir Roushad, Su Ping oon, and Neomy Storch. 2015. What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, 28:39–52.

Ute Knoch, Amir Roushad, and Neomy Storch. 2014. Does the writing of undergraduate ESL students develop after one year of study in an english-medium university? *Assessing Writing*, 21:1–17.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

Diane Larsen-Freeman. 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4):590–619.

Rémi Lavalley, Kay Berkling, and Sebastian Stüker. 2015. Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

Ildikó Pilán and Elena Volodina. 2016. Classification of language proficiency levels in swedish learners' texts. In *Proceedings of Swedish language technology conference*.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

Robert Reynolds. 2016. *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications*. Ph.D. thesis, UiT - The Arctic University of Norway.

Peter Robinson. 2001. Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, 22(1):27–57.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka.

Peter Skehan. 1996. A framework for the implementation of task-based instruction. *Applied Linguistics*, 17(1):38.

Eric Velleman and Thea van der Geest. 2014. Online test tool to determine the cefr reading comprehension level of text. *Procedia computer science*, 27:350–358.

Mary Lou Vercellotti. 2015. The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 38(1):90–111.

Marjolijn Verspoor, Monika S. Schmid, and Xiaoyan Xu. 2012. A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21(3):239–263.

Nina Vyatkina. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, 96(4):576–598.

Zarah Weiss. 2017. Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects. Master's thesis, University of Tübingen, Germany.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA.

Zarah Weiss and Detmar Meurers. in press. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers. 2019. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Honolulu.

Weiwei Yang, Xiaofei Lu, and Sara Cushing Weigle. 2015. Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28:53–67.

Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Hyung-Jo Yoon. 2017. Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66:130–141.

Hyung-Jo Yoon and Charlene Polio. 2016. The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, pages 275–301.

A Appendices

↓Obs/Pred→	1/2	3/4	5/6	7/8	$\sum$
1/2	**1217**	813	0	0	2030
3/4	430	**4810**	0	0	5240
5/6	0	0	**3029**	1591	4620
7/8	0	0	1590	**2850**	4440
$\sum$	1647	5623	4619	4441	16330

Table 10: Confusion matrix for the best model in study 1 (all feat. + meta) summed across iterations

↓Obs/Pred→	1/2	3/4	$\sum$
1/2	**1232**	798	2030
3/4	449	**4791**	5240
$\sum$	1681	5589	7270

Table 11: Confusion matrix for best elementary school model in study 2 (all feat. + meta) summed across iterations

↓Obs/Pred→	5/6	7/8	$\sum$
5/6	**3049**	1571	4620
7/8	1497	**2943**	4440
$\sum$	4546	4514	9060

Table 12: Confusion matrix for best secondary school model in study 2 (all feat. + meta) summed across iterations

↓Obs/Pred→	1/2	3/4	$\sum$
1/2	**51**	39	90
3/4	9	**162**	171
$\sum$	60	201	261

Table 13: Confusion matrix for the best model for elementary school in study 3 (Error rate)

↓Obs/Pred→	5/6	7/8	$\sum$
5/6	**91**	39	130
7/8	51	**60**	111
$\sum$	142	99	241

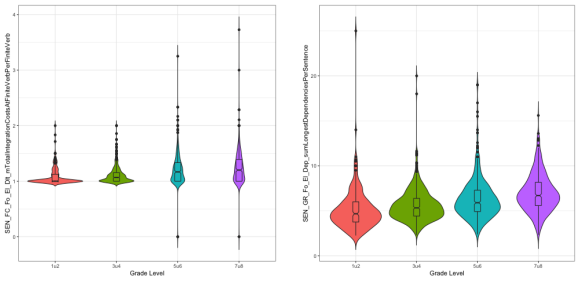
Table 14: Confusion matrix for the best model for secondary school in study 3 (all feat. + meta)

Feature name	Set	Merit
Overall errors / W	Error rate	.166
Root type-token ratio	Lexical	.150
Corrected type-token ratio	Lexical	.150
Number of words	Clausal	.137
Capitalization errors / W	Error rate	.128
HDD	Lexical	.124
Corrected verb variation	Lexical	.110
Squared verb variation	Lexical	.110
Word splitting + hyphenation errors / W	Error rate	.108
P(Subject→Nothing)	Discourse	.106
P(Nothing→Nothing)	Discourse	.104
P(Nothing→Subject)	Discourse	.099
Number of sentences	Clausal	.094
P(Nothing→Object)	Discourse	.093
Yule's K	Lexical	.091
MTLD	Lexical	.088

Table 15: Top features in information gain ranking for grade-level distinction in elementary school

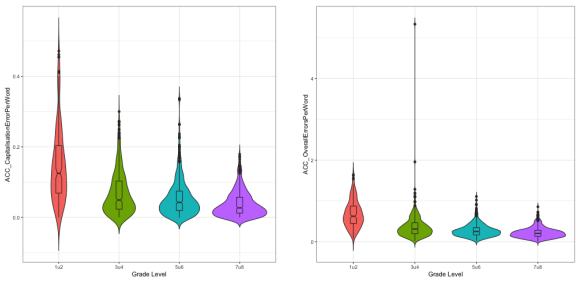
Feature name	Set	Merit
Yule's K	Lexical	.030
Capitalization errors / W	Error rate	.029
($\sum$ max. dep.) / S	Human processing	.026
MTLD	Lexical	.023
Verbs / t-unit	Clausal	.023
Verbs / S	Clausal	.023
HDD	Lexical	.022
Overall errors / W	Error rate	.022
Nouns / W	Lexical	.021
$\sum$ Non-terminal nodes / tree	Clausal	.021
W / S	Clausal	.021
to infinitives / S	Lexical	.020
Uber index	Lexical	.020
Temporal connectives / S	Discourse	.019
$\sum$ Non-terminal nodes / W	Clausal	.019
Clauses / S	Clausal	.017

Table 16: Top features in information gain ranking for grade-level distinction in secondary school



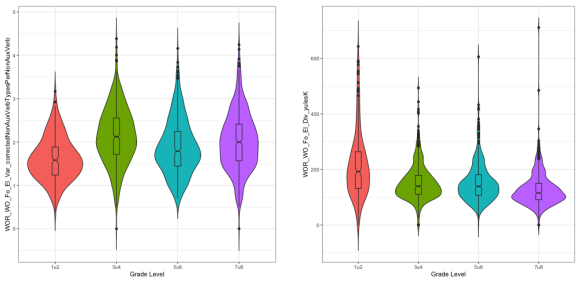
(a) DLT integration cost (m) (b) Max. dependency / S

Figure 1: Most informative human processing features



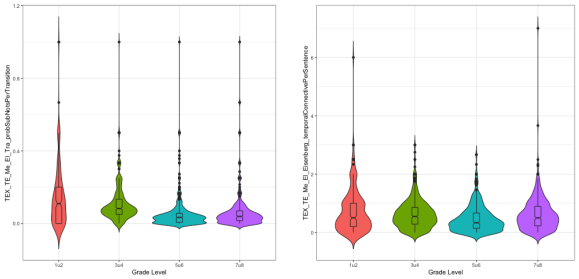
(a) Capitalization errors (b) Overall errors

Figure 2: Most informative error rate features



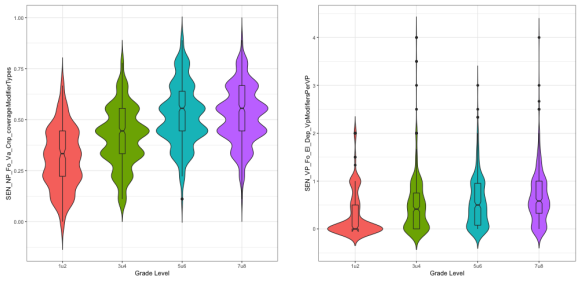
(a) Corrected verb variation (b) Yule's K

Figure 3: Most informative lexical features



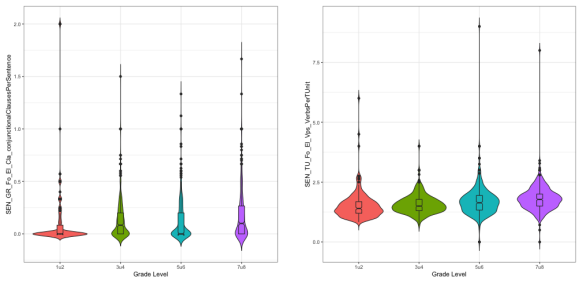
(a) Subject transitions (b) Temporal connectives

Figure 4: Most informative discourse features.



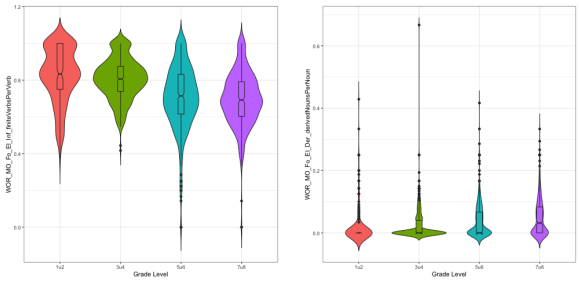
(a) NP modifier coverage (b) Verb modifiers / VP

Figure 5: Most informative phrasal features.



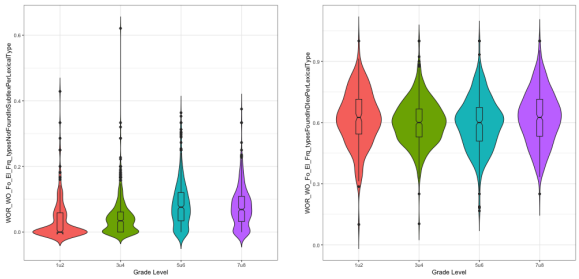
(a) Conjunction clauses / S (b) Verbs / t-unit

Figure 6: Most informative clausal features.



(a) Finite verbs / verb (b) Derived nouns / noun

Figure 7: Most informative morphology features.



(a) Words in Subtlex-DE (b) Words in dlexDB

Figure 8: Most informative language use features

Analyzing the linguistic complexity of German learner language in a reading comprehension task

Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality

Zarah Weiss and Detmar Meurers
University of Tübingen

While traditionally linguistic complexity analysis of learner language is mostly based on essays, there is increasing interest in other task types. This is crucial for obtaining a broader empirical basis for characterizing language proficiency and highlights the need to advance our understanding of how task and learner properties interact in shaping the linguistic complexity of learner productions. It also makes it important to determine which complexity measures generalize well across which tasks.

In this paper, we investigate the linguistic complexity of answers to reading comprehension questions written by foreign language learners of German at the college level. Analyzing the corpus with computational linguistic methods identifying a wide range of complexity features, we explore which linguistic complexity analyses can successfully be performed for such short answers, how learner proficiency impacts the results, how generalizable they are across different contexts, and how the quality of the underlying analysis impacts the results.

Keywords: complexity analysis, L2 German, proficiency assessment, reading comprehension, natural language processing

1. Introduction

Complexity research is an established area in Second Language Acquisition (SLA), where Complexity together with Accuracy and Fluency form the core dimensions of second language (L2) performance and proficiency known as the CAF triad

https://doi.org/10.1075/ijlcr.20006.wei
International Journal of Learner Corpus Research 7:1 (2021), pp. 83–130. issn 2215-1478 | e-issn 2215-1486

(Housen & Kuiken, 2009). As empirical basis for the research of CAF, the studies traditionally are based on longer written learner productions, such as essays collected in learner corpora. More recently, there is increasing interest in different task types, which have been shown to heavily influence CAF constructs (Alexopoulou, Michel, Murakami, & Meurers, 2017; Biber, Gray, & Staples, 2016; Caines & Buttery, 2017). Since differences in both – language proficiency and tasks – can cause systematic variation in CAF constructs, research needs to cleanly separate proficiency effects from task effects. This is crucial for supporting valid analyses of interlanguage (Meurers & Dickinson, 2017) and obtaining a characterization of language proficiency grounded in SLA (Tracy-Ventura & Myles, 2015).

In terms of the complexity measures targeted, research has mostly focused on measures of lexical and syntactic complexity and primarily analyzed English. This limited research scope has been criticized as overly reductionist (Housen, De Clercq, Kuiken, & Vedder, 2019) and other linguistic domains and different L2s have started to attract more attention. To compute a wider range of linguistic complexity measures for substantial samples of learner data, research increasingly makes use of computational linguistic methods to automate the analyses on which complexity measures are based. Since the underlying Natural Language Processing (NLP) models are generally trained on native language data, the validity of the analysis cannot be taken for granted for learner language and can involve substantial conceptual challenges (Meurers, 2020; Meurers & Dickinson, 2017). It thus is important to determine how much the computation of complexity features is impacted by the effect that learner language characteristics have on the automated NLP analyses.

In this paper, we bring these strands together by (a) extending the range of task types used in complexity research to include short answers given in response to reading comprehension questions, (b) computing a broad range of complexity measures for German and exploring the generalizability of results across task contexts varying to different degrees, and (c) testing the impact of learner language characteristics on the automated complexity analysis.

2. Related work

A current overview of SLA research on linguistic complexity is available through the recent special issue on the topic (Housen et al., 2019), so we focus the discussion here on the research strands introduced above that motivate the current paper. Starting with the relevance of analyzing language from a range of contexts and of considering the nature of the task when analyzing the learner language that was produced for it, task effects on linguistic complexity are receiving increasing attention.

This includes work grounded in the Task-based Language Learning approach to SLA (Alexopoulou et al., 2017; Michel, Murakami, Alexopoulou, & Meurers, 2019), as well as studies on the influence of register (Biber et al., 2016), genre (Yoon & Polio, 2016), and topic (Yoon, 2017). This research shows that task differences can introduce systematic variation of CAF and that such task effects have to be controlled when using CAF to characterize language proficiency. While learner corpus construction so far has mostly focused on learner essays, some tools have been designed to support teachers in collecting rich task-based data (e.g., WELCOME for reading comprehension tasks, Ott, Ziai, & Meurers, 2012) and some studies highlight the importance of task design for making learner corpora relevant for SLA research (e,g., Tracy-Ventura & Myles, 2015). Recent computer-based language use contexts also provide access to task information, such as task-based L2 text chat data (Ziegler, 2018), and data elicited in computer-based learning environments offering a broad range of activities, such as EFCamDat (Geertzen, Alexopoulou, & Korhonen, 2013), arguably can play an important role in helping overcome the limitations of traditional learner corpora.

Turning to the nature of the complexity measures being analyzed, morphological complexity is increasingly receiving attention, also due to a broadening of the set of languages being analyzed to include more morphologically rich languages. Brezina and Pallotti (2019) propose a new measure of morphological complexity that correlates with Italian L2 proficiency and distinguishes between native and advanced non-native writing in Italian but not in English. De Clercq and Housen (2019) report similar findings for morphological complexity measures on spoken L2 French compared to English.

The endeavor of broadening the scope of complexity research can also build on analytic methods that have been developed for readability assessment for a broader range of languages, including French (François & Fairon, 2012), German (Hancke, Vajjala, & Meurers, 2012; Weiss & Meurers, 2018), Swedish (Pilán, Vajjala, & Volodina, 2015), and Italian (Dell'Orletta, Montemagni, & Venturi, 2014). The close connection between linguistic complexity analysis in SLA and readability research is emphasized by Vajjala and Meurers (2012), demonstrating the successful use of SLA complexity measures for predicting text readability. In current readability research, the extraction of a range of complexity measures is generally combined with machine learning techniques to explicitly model and predict text readability (e.g., Crossley, Skalicky, & Dascalu, 2019; Weiss & Meurers, 2018). In a similar vein, complexity measures are also used to characterize and predict academic writing development (e.g., Crossley, Weston, Sullivan, & McNamara, 2011; Staples, Egbert, Biber, & Gray, 2016; Weiss & Meurers, 2019a). These different application domains make use of a range of measures well beyond traditional lexical and syntactic complexity. In research on academic writing quality, discourse com-

plexity and text cohesion have played a major role (cf. Crossley, 2020 and references therein). Psycholinguistic measures focused on human processing demands at the sentence level have successfully been adapted to the tasks of readability and proficiency assessment (Shain, van Schijndel, Futrell, Gibson, & Schuler, 2016; Weiss & Meurers, 2019b). The integration of measures from diverse subfields to broaden the scope of complexity research has been shown to substantially increase accuracy and robustness of the models (Crossley, 2020; Weiss & Meurers, 2019b).

3. This study

In this paper, we want to broaden the empirical base of complexity research and investigate a number of factors: the feasibility of reliably analyzing linguistic complexity based on short learner responses, the generalizability of the results, and the quality of the NLP involved in automating the analysis. Our exploration is based on a corpus of short answers to reading comprehension questions written by college-age learners of German in the US, the CREG corpus (Ott et al., 2012). Reading comprehension activities are very commonly used and well-suited to foster language learning in the way they require learners to interact with both form and meaning in input and output. But as far as we are aware, answers to reading comprehension questions have so far not been studied in linguistic complexity research. Compared to essays, the relatively short answers pose a challenge in providing less language material that can be analyzed, which can negatively impact the robustness of complexity measures and the range of complexity measures that can be computed. For example, computing lexical diversity or averages of morphological, lexical, or syntactic measures becomes more robust for longer texts, and some measures of discourse cannot be applied to short answers. The explicit task context provided by the given reading text and the comprehension question also defines a precise functional setting delineating the range of language forms that can be used to answer the question – in line with variationist linguistics specifying variables to determine the possible variants that can be studied and interpreted (Tagliamonte, 2011) and register research emphasizing the need to analyze language that is functionally associated with a situational context (Biber et al., 2016: 643).

Against this background, we pursue the following research questions (RQ):

1. Can we model L2 proficiency using broad linguistic complexity analysis on the limited evidence provided by short answers to reading comprehension questions?
2. To what extent is such a model generalizable across task contexts provided by different questions and reading texts?

3. How much do the characteristics of learner language impact the NLP-based extraction of complexity measures and the overall complexity model?

To answer these questions, we first introduce the CREG corpus and the data sets used in the studies (Section 4) as well as the complexity analysis pipeline we employed (Section 5). We then spell out the machine learning experiments designed to address the first two research questions in Section 6, which includes a visualization of the experimental set-up and analysis. Building on the concepts and methods introduced in that section, we then address the third research question in Section 7, reporting on the performance of the analysis pipeline on learner language when compared to a manual analysis, before concluding the paper in Section 9.

4. Data

Our analyses are based on the *Corpus of Reading Comprehension of German* (CREG; Ott et al., 2012; Ziai, 2018). CREG is a task-based corpus of reading exercises that supports the analysis of learner language in the explicitly given task context used in eliciting the data. The corpus consists of data elicited over the course of four years in beginner and intermediate level German courses taught in two German language programs in the US, at the University of Kansas (KU) and the Ohio State University (OSU). Together with the learner language in the answers, the corpus contains explicit meta-information about text, questions, target answers used to coordinate expectations across classes, and teacher ratings of the learner answers in terms of whether they answered the question. The students almost all specified English as their L1 (97.8%), with very few reporting having German-speaking parents (5.3%) or having been to Germany for more than three months (13.6%).

As illustrated in Figure 1, a reading exercise in CREG consists of one reading text, one or more reading comprehension questions, one or more possible target answers defined by the programs' language teachers, and transcriptions of student answers to each question. Student answers were originally hand-written and were all transcribed by two annotators using the *WEb-based Learner COrpus Machine* (WELCOME, Ott et al., 2012). While in most cases the transcriptions are identical, for some answers there are differences, e.g., due to one transcriber not having transcribed all of the answer. For the purposes of this article, we always chose the longest transcription.

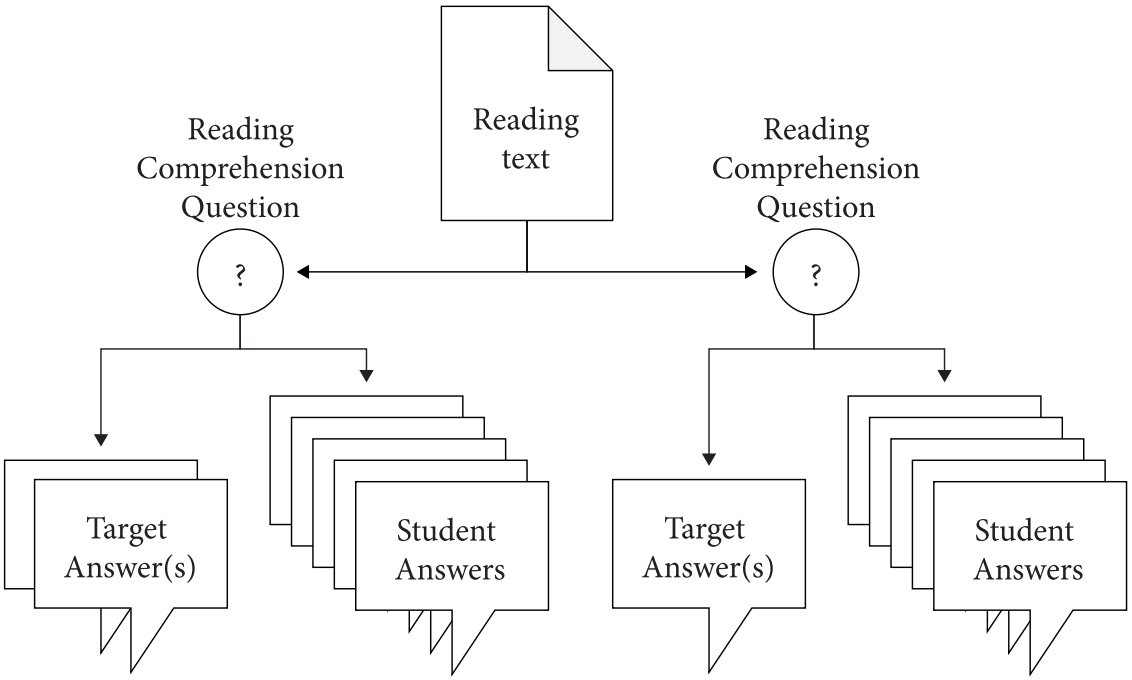


Figure 1. Components of a reading exercise in CREG

4.1 CREG-29K

We focus the analysis in this paper on the four courses at the beginner-level at both universities, corresponding to CEFR levels A1 to A2. These levels are the target proficiency levels at the end of the course as specified by the program directors. For the studies in this article, we use this course level as approximation of student proficiency. Despite placement testing, students in the same course naturally differ in ability and individually develop during the course, so the proficiency labeling is relatively noisy.

The raw corpus we used consists of 29,019 student answers. We removed 106 student answers for which manual linguistic annotations are available (see Section 4.3) in order to keep those as a separate evaluation data set for the final study (Section 7). The remaining 28,913 student answers constitute our CREG-29K corpus. On average students contributed 16.41 answers ($SD = 15.00$) at their course level.

4.2 CREG-KU, CREG-OSU, and CREG-7K

The distribution of student answers across course levels is highly imbalanced. To answer our first two research questions, we therefore extracted three sub-corpora from CREG-29K that are balanced across course levels using stratified random sampling: CREG-7K contains 6,548 student answers elicited in equal parts at KU and OSU (7,432 sentences; 46,987 words). Table 1 spells out the overall number of student answers, sentences, and words by course level (indicated using the corresponding target CEFR level) and institution (KU and OSU).

We also created two corpora divided by institution to investigate differences between KU and OSU: CREG-KU contains 7,839 student answers elicited at KU (8,472 sentences; 44,968 words). CREG-OSU contains 3,259 student answers elicited at OSU (3,865 sentences; 27,489 words). Table 2 shows the number of student answers, sentences, and words by course level for both sub-corpora.

Table 1. Corpus statistics of CREG-7K across course levels and universities

	A1.1		A1.2		A2.1		A2.2	
	KU	OSU	KU	OSU	KU	OSU	KU	OSU
#answers	742	733	901	905	821	815	814	817
#sentences	794	771	938	1,032	929	988	891	1,089
#words	2,697	4,315	4,128	6,560	5,773	7,748	6,688	9,078

Table 2. Corpus statistics of CREG-KU and CREG-OSU across course levels

	CREG-KU				CREG-OSU			
	A1.1	A1.2	A2.1	A2.2	A1.1	A1.2	A2.1	A2.2
#answers	1,995	1,901	1,977	1,966	736	905	809	810
#sentences	2,100	1,982	2,263	2,127	773	1,032	980	1,080
#words	7,305	8,384	13,625	15,654	4,323	6,560	7,639	8,967

4.3 CREG-104

Ott and Ziai (2010) manually created a dependency annotation for 106 KU student answers for course levels A1.1, A2.1, and A2.2 using three trained annotators and a final arbitration step to resolve conflicts. For our third study, we converted the dependency annotation to the scheme of S. Brants, Dipper, Hansen, Lezius, and Smith (2002), used by the NLP tools discussed in Section 5, by mapping dependency labels and adjusting the sentence segmentation so that every dependency graph has a single root. We also excluded two student answers that were written entirely in English, leaving 104 student answers with 780 words in 110 sentences. Table 3 provides the details for this CREG-104 corpus.

Table 3. Corpus statistics of CREG-104 across course levels

	A1.1	A1.2	A2.1	A2.2
#answers	25	0	32	47
#sentences	25	0	33	52
#words	165	0	221	394

We augmented the CREG-104 annotation with a morphological analysis. For this, we manually corrected the output of the morphological analysis of the Mate tools (for details see Section 5.2). We thereby obtain annotations for lemmas as well as case (nominative, accusative, genitive, dative), number (singular, plural), gender (male, female, neuter), person (first, second, third), tense (simple past, simple present), verb mode (indicative, imperative, subjunctive), and degree of comparison (positive, comparative, superlative). We use these manual annotations as a reference to evaluate the purely automatic NLP analyses in Section 7. An example for the manual reference annotation of a student answer in CREG-104 is provided in Figure 2.

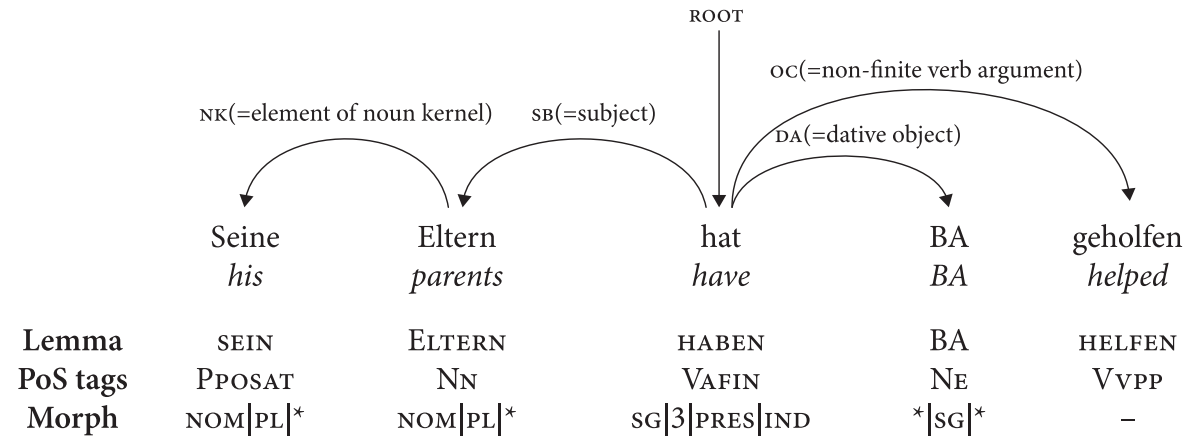


Figure 2. Student answer (word-by-word translation in italics) with manual reference annotation of dependencies, lemmas, parts-of-speech tags, and morphological features (here: CASE|NUMBER|GENDER for nouns and NUMBER|PERSON|TENSE|MODE for finite verbs)

4.3.1 *Manual annotation of learner language and target hypotheses*

Since linguistic categories and annotation schemes were generally developed for the analysis of well-formed native language, analyzing learner language poses substantial challenges (Meurers, 2015). For example, when analyzing learner language, the distributional, morphological, and lemma evidence for determining parts-of-speech (PoS) often fails to converge on a single category (Díaz-Negrillo, Meurers, Valera, & Wunsch, 2010), and the bottom-up form-based characteristics often do not line up with the top-down meaning-based properties (Meurers & Dickinson, 2017). In order to obtain a manual linguistic analysis for various levels of linguistic annotation for learner language, it is thus essential to explicitly define the target hypothesis that serves as the basis of the respective linguistic annotation (Lüdeling, 2008). Depending on the analysis purpose, multiple layers of annotation are required, which can be based on distinct target hypotheses (Lüdeling, Walter, Kroymann, & Adolphs, 2005).

This can be illustrated with the student answer from CREG-104 shown in Figure 2. The surface form of each individual expression adheres to the German norm in the sense that each of those word forms exists in German. This supports a straightforward local surface form analysis. But at the clausal level, the syntactic case marking and the subject verb agreement relations of the finite verb *hat* are incompatible with the nominal forms realized in the sentence.[1] If *seine Eltern* (his parents) is the subject, the plural verb form would be required (***haben***). So the target hypothesis for that analysis would be *Seine Eltern haben BA geholfen*, which means that someone named BA was helped by his parents. Alternatively, BA could also be the nominative singular subject of the sentence, but then *seine Eltern* would need to be realized in accusative case. So the target hypothesis for that analysis is ***Seinen*** *Eltern hat BA geholfen* and the interpretation of that sentence becomes that BA helped his parents. Given that these two analyses differ in their interpretation, the explicit task context helps us decide which of the analyses to chose, i.e., which morphological evidence not to take at face value. This sentence was written to answer the question "Who did BA help?" so based on this top-down task information, we can determine that the analysis with BA as the subject is the right one in this context.

Stepping back from the example, in general we can establish two different target hypotheses, a local form-based target hypothesis (TH1) and a meaning-based target hypothesis (TH2). We cannot base our manual reference annotations solely on TH1, because it does not allow us to conduct a dependency analysis of entire clauses. TH2, on the other hand, does not make full use of the morphological evidence. If one simply glossed over the missing accusative case marking, such a robustly meaning-focused analysis would miss out on information that is potentially very relevant for characterizing language acquisition, the incremental acquisition of morpho-syntax. Due to the diverging requirements that these linguistic reference annotations need to satisfy, we use both target hypotheses and make them explicit in the reference annotation. We follow a local form-based target hypothesis for our PoS, lemma, and morphological reference annotations, while using a meaning-based target hypothesis for the dependency annotation.

1. We here are discussing morpho-syntactic dependencies. At the semantic level, the main lexical predicate *helfen* (help) would be the head.

5. Automatic complexity analysis

To analyze linguistic complexity, we calculate 297 features of linguistic complexity using our complexity code (Weiss & Meurers, 2018), which has been successfully used to model the complexity of longer text productions for German L2 proficiency assessment (Weiss & Meurers, 2019b), early academic language development (Weiss & Meurers, 2019a), and readability assessment (Weiss & Meurers, 2018). In the following, we first describe the complexity features we calculated (Section 5.1) before elaborating on the NLP pipeline used for this (Section 5.2).

5.1 Feature description

Linguistic complexity is the "degree to which language is elaborate and varied" (R. Ellis, 2003: 340). In terms of the taxonomy of Housen, Kuiken, and Vedder (2012), we focus on the absolute linguistic complexity in the lexical, morphological, phrasal, clausal, and discourse domains. The code also extracts features of relative complexity that are motivated by usage-based theories of language learning (e.g., N. C. Ellis, 2002) grouped under language use, as well as insights from psycholinguistic research included under the label human processing. We briefly characterize here the types of features we calculate for each domain. A comprehensive list of all features, their definitions, and definitions of the linguistic units used to measure them is included in Appendices A and B.

Lexical complexity

We calculate 34 measures of lexical complexity. These predominantly include long-established measures of lexical diversity and variation, including different types of general or PoS-specific type-token ratios as well as the measure of textual lexical diversity (MTLD; McCarthy, 2005). This category also includes other PoS-based ratios and some measures of semantic relatedness and specificity.[2]

Morphological complexity

We include 41 measures of morphological complexity that cover the domains of inflection, derivation, and compounding. These include features of the elaboration of compounds such as average noun compound depth, measures of the expression of case or tense, or noun derivation measures.

---

2. While traditionally lexical sophistication is also grouped under lexical complexity, we include it under "language use" below to maintain a clear distinction between features of absolute and relative complexity.

Phrasal complexity

We compute 47 measures of phrasal complexity that assess the elaboration as well as variation of phrasal modification. Aside from more coarse-grained measures such as average noun phrase length, this also includes specialized measures for noun and verb phrase complexity as well as measures of the variation of phrase modification. We include verb phrase modification here instead of under clausal complexity since the verb clusters of German offer substantial possibility for complexification below the clausal level. Likewise, the category also includes several measures of elaborate grammatical constructions such as periphrastic tense patterns.

Clausal complexity

We analyze 25 measures of clausal complexity, focusing predominantly on clausal subordination and coordination. Aside from more global measures of overall clausal complexity such as t-units per sentence, this also includes more fine-grained measures of clausal elaboration such as relative clauses per clause.

Discourse complexity

We identify 64 measures of discourse complexity and cohesion. These include explicit cohesion measures based on connectives, and implicit cohesion measures such as grammatical transitional probabilities from one sentence to another, e.g., the probability of a subject being repeated in the next sentence as an object, and global as well as local argument, stem, and content overlap measures. Finally, this category also includes measures of pronoun use.

Language use

We compute 58 features of language use in terms of lexical sophistication measures (word frequencies) and age of active use measures. The word frequency measures are based on several frequency data-bases including news data, captions of movies, books, and childrens' writings.

Human processing

We include 25 features of human processing based on dependency lengths and the Dependency Locality Theory (DLT; Gibson, 2000) for human sentence processing. The theory bases human processing costs on the locality of dependents and the cost of storing incomplete discourse structures. Its predictions are used to assess the average and maximal integration cost at the finite verb, using several DLT parametrizations suggested by Shain et al. (2016).

Surface measures

Additionally, the code calculates three surface text features: number of words, number of sentences, number of paragraphs.

5.2 System description

The code is written in Java and extracts complexity measures from plain text input in a three-step process. First, it generates automatic linguistic annotations at various levels using an elaborate NLP pipeline, including:

- Sentence segmentation and tokenization with *Apache OpenNLP 1.9.1* with their model trained on the Leipzig corpus (Goldhahn, Eckart, & Quasthoff, 2012).[3]
- PoS tagging, lemmatization, morphological analysis, and dependency parsing using the Mate tools 1.3 (Björkelund, Bohnet, Hafdell, & Nugues, 2010; Bohnet & Nivre, 2012) with their model trained on the dependency conversion of the Tiger treebank (S. Brants et al., 2002) described in Seeker and Kuhn (2012) without ellipses.
- Constituency parsing using the Stanford parser from the *CoreNLP 3.9.2* pipeline (Chen & Manning, 2014) with their model trained on the Negra corpus (T. Brants, Skut, & Uszkoreit, 1999). To ensure consistency in the linguistic annotation, this step is based on the PoS tags produced by the Mate tools.
- Topological field parsing using the Berkeley parser 1.7 (Petrov & Klein, 2007) with the model by Ziai (2018) trained on the TüBa-D/Z treebank (Telljohann, Hinrichs, & Kübler, 2004).
- Compound splitting using the dictionary-based *jWordSplitter 3.4*.[4]

These models provide state-of-the-art performance on standard German data (cf. references cited). For an estimation of their performance on learner data, please see Section 7.

Following the linguistic annotation step, the code extracts all information required for the calculation of the complexity features. This predominantly involves counting linguistic constructions, but also looking up word frequencies in frequency data-bases or semantic relations in *GermaNet 11.0* (Hamp & Feldweg, 1997).

Finally, the code calculates the complexity features based on the extracted counts. Each analysis step produces its own intermediate output, which serves as input for the following step. As such, the extraction of linguistic constructions and complexity features can also be partially or fully based on external analyses such as manual reference annotations.

3. https://opennlp.apache.org/
4. http://www.danielnaber.de/jwordsplitter/

6. Determining German L2 proficiency using linguistic complexity analysis

To address the first research question regarding the extent to which it is possible to determine German L2 proficiency based on the limited evidence provided by short answers to reading comprehension questions, we build and compare course-level classifiers. We use data from CREG-KU, CREG-OSU, and CREG-7K to test to what extent it is possible to predict the four course levels (A1.1, A1.2, A2.1, and A2.2) based on the linguistic complexity analysis of the student answers (Section 6.1). In Section 6.2, we then conduct a more elaborate exploration to see how well our classifiers generalize across task contexts that vary to different degrees.

6.1 Course-level classification

6.1.1 *Set-up of study 1*

We analyzed the student answers in the CREG corpus to extract 297 complexity features using the approach described in Section 5. We then removed all features that showed little variability on this data set, receiving the same value for more than 80% of the data points. This applied to 150 complexity features leaving 147 features for the analysis (marked by superscript a in Appendix B). We calculated the z-scores for each complexity feature and split the data into the CREG-KU, CREG-OSU, and CREG-7K sub-corpora introduced in Section 4.2. We then removed extreme outliers.[5] We obtained training, development, and test sets for each sub-corpus using a 70/20/10 split stratified by course level. After obtaining the splits, we trained an ordinal random forest (ORF) classifier[6] for which we performed hyper-parameter tuning[7] on the development set. Figure 3 provides a visual overview of the approach.[8]

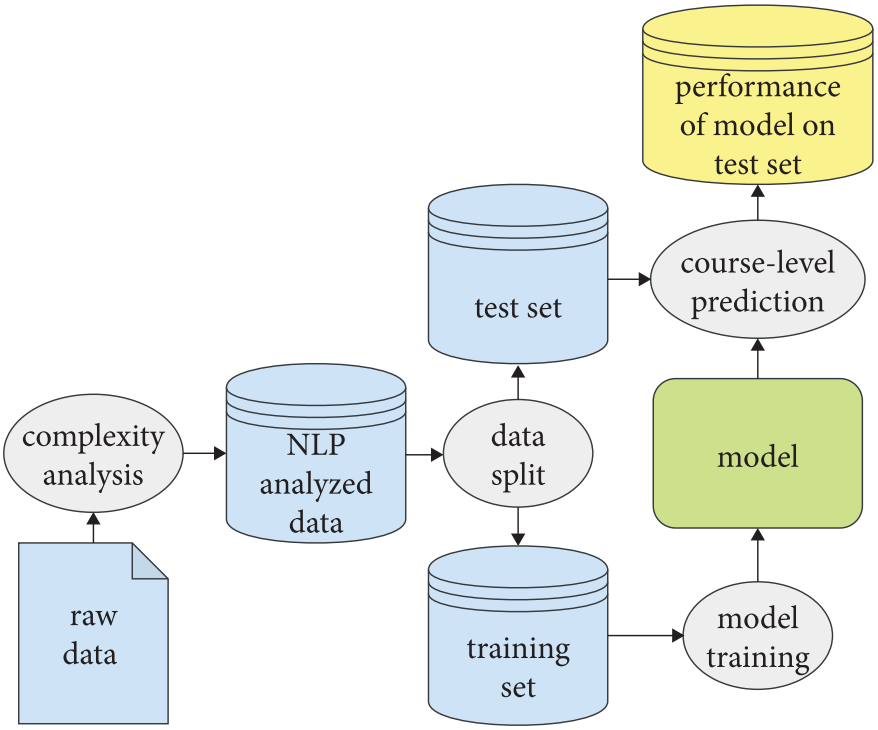
---

5. To not interfere with the highly variable nature of the data, we proceeded very conservatively, using nine standard deviations from the mean as a threshold.

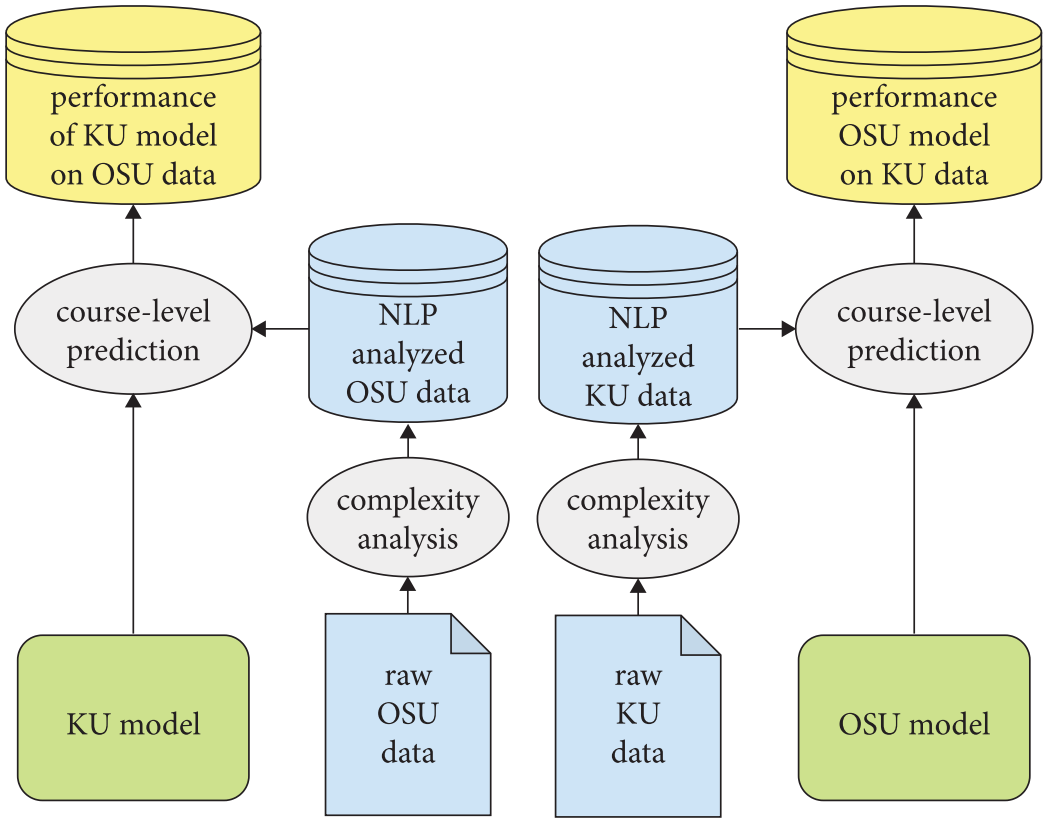
6. An ordinal random forest is a random forest that treats the predicted variable as ordinal rather than nominal data. We initially also explored using a Support Vector Machine using polynomial or radial kernels. Since ORF systematically performed best, we only report the ORF results. Since these machine learning algorithms are relatively robust against collinear features, the correlation between complexity measures was no particular concern for these experiments.

7. Hyper-parameter tuning is a process that sets the parameters of the machine learning algorithm used.

8. To keep things readable, the figures do not show the separate development set, which is like a test set but used for model tuning during development, as is usual in machine learning.



a. Train/test procedure



b. Cross-corpus test procedure

Figure 3. Set-up for RQ1: For each corpus (CREG-7K/OSU/KU), student answers are automatically analyzed and split into distinct sets to train and test the L2 proficiency model (3a). The KU and OSU models generated by this procedure are additionally cross-corpus tested on the respective other corpus (3b)

6.1.2 *Results of study 1*

Table 4 shows the results of the machine learning experiments for the first study. For each classifier, it reports the course-level prediction accuracy on the regular test set (KU/OSU/$7K_{test}$) and for the models trained on the KU or OSU training data it also reports the performance for the respective other full corpus (OSU or KU) for cross-corpus testing. For each accuracy result, we analyzed whether it significantly improves on the majority baseline.[9]

We see that the classifier trained on the KU training set performs well on the KU test set with an accuracy of 84.21%. The model trained on the OSU training set achieves an accuracy of 81.45% on the OSU test data. So the rich feature set successfully identifies differences in linguistic complexity between course levels for data collected in independent language programs.

Table 4. Classifier performance of course-level prediction across sub-corpora

Trained on	Tested on	Acc.	95% CI	Baseline	P-Value
CREG-KU_{train}	CREG-KU_{test}	84.21	[82.67; 85.66]	25.45	$< 2^{-16}$
	CREG-OSU	28.35	[26.81; 29.93]	27.77	0.2342
CREG-OSU_{train}	CREG-OSU_{test}	81.45	[78.87; 83.85]	27.77	$< 2.2^{-16}$
	CREG-KU	28.83	[27.83; 29.85]	25.45	6.559^{-12}
CREG-$7K_{train}$	CREG-$7K_{test}$	73.58	[72.24; 74.88]	26.95	$< 2^{-16}$

For the trained models, where the evidence contributed by the different features is weighted based on the training data, we find that the models do not generalize across universities. When tested on the CREG-KU corpus, the accuracy of the classifier trained on CREG-OSU_{train} drops to 28.83%, barely outperforming the majority baseline, and the classifier trained on CREG-KU_{train} does not perform significantly above the majority baseline when tested on the CREG-OSU corpus.

For the classifier trained on CREG-$7K_{train}$, which contains KU and OSU data in equal parts, the accuracy of 73.57% shows that student answers elicited at KU and those elicited at OSU do share some linguistic complexity characteristics allowing the classifier to generalize from one data set to the other. So the fact that the models trained on the CREG-KU/CREG-OSU corpora did not generalize to the other university's test corpus has to be due to other factors, which we investigate further

9. Accuracy is the number of correctly classified items divided by all items (Tharwat, 2018) and will be used as a performance measure throughout this article. For significance testing, we used one-sided t-tests with $H_1 = Acc. > Baseline$ based on the confusion matrices, as implemented by the R package *caret*.

in the second study in Section 6.2. Another interesting research question, which for space reasons is beyond the scope of this paper, would be to explore which features work best across universities and for which types of subtasks (e.g., question types).

6.2 Generalizability of complexity modeling

6.2.1 *Set-up of study 2*

To go beyond basic cross-corpus evaluation and gain insights into what determines generalizability of the results, we need to systematically investigate under which conditions the linguistic complexity models generalize. We therefore perform machine learning experiments that differ in terms of how many characteristics of the reading task are shared between the training and the test data. This is illustrated in Table 5. None of the test sets contain answers that are part of the training data. But the settings differ in terms of whether data for the same questions, reading texts, or from the same university are included in the training and test sets.

Table 5. Overview of the characteristics shared between training and test data

	Answer	**Question**	**Text**	**University**
Regular test set		✓	✓	✓
Held-out question set			✓	✓
Held-out text set				✓
Cross-corpus set				

In the first study, we had used a regular test set (containing randomly selected answers to questions for which possibly other answers were part of the training set) and a cross-corpus test set (containing answers elicited at a different university for different tasks). We now introduce two additional test sets standing in between the previous two in terms of how many characteristics they share with the training set. In machine learning, we speak of held-out data to indicate that certain data was not used in training a classifier in order to observe how well this classifier generalizes to the held-out type of data. The held-out question test set (OSU/KU/7K$_{\text{HoQ}}$) includes answers to questions that are not included in the training set (row 2 in Table 5). The held-out text test set (OSU/KU/7K$_{\text{HoT}}$) includes answers to questions that relate to texts that did not occur in the training data (row 3 in Table 5).

To obtain the data subsets for our second study, we proceeded as follows: We first split each of the overall data sets (OSU, KU or 7K) into a 10% test data set (HoQ-HoTest and HoT-HoTest) satisfying the held-out criterion (questions or texts). From the remaining 90% we randomly sampled an ordinary 10% test set

(HoQ-RTest, HoT-RTest) and a 10% development set (HoQ-Dev, HoT-Dev), and the remaining 70% served as training set (HoQ-train, HoT-train). So overall, we obtain two data set splits for each of the OSU, KU, and 7K subcorpora: one 10/10/10/70 split for the held-out question condition and one such split for the held-out text condition. Apart from the two new data splits used, the set-up is identical to that of Study 1 in Section 6.1.1. Figure 4 provides a visual overview of the approach.[10]

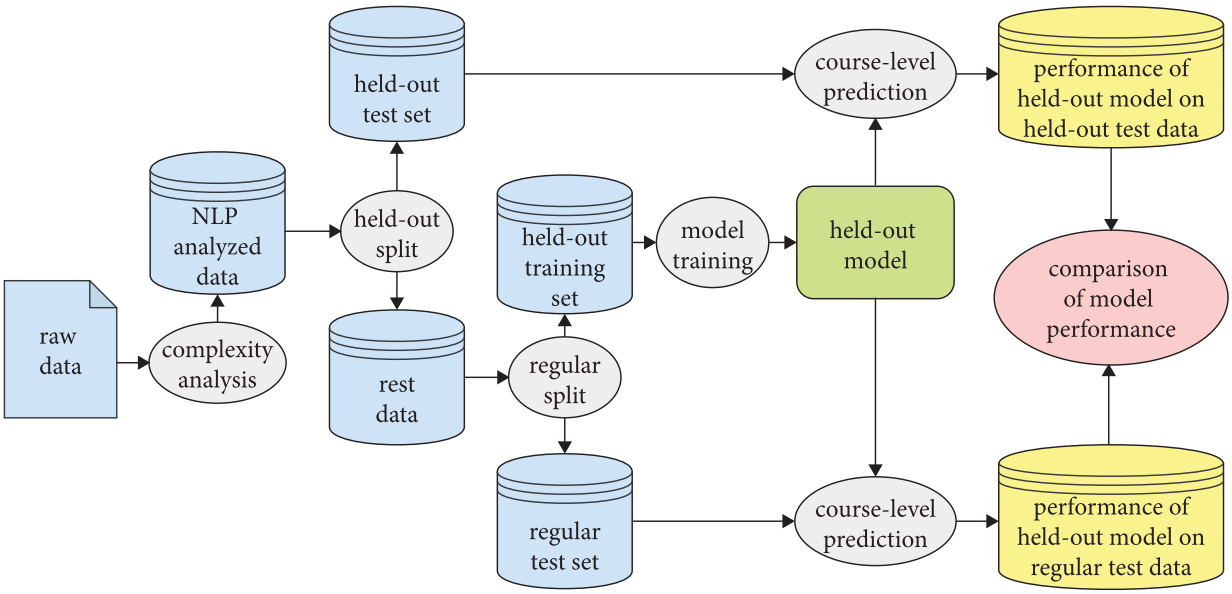


Figure 4. Set-up for RQ2: Corpus data is analyzed and split into training and test data, once ensuring held-out questions and once ensuring held-out texts

6.2.2 *Results of study 2*

Table 6 shows the results of our machine learning experiment using held-out reading comprehension questions and Table 7 the results using held-out reading texts. They report the classification performance in terms of accuracy on the regular test set and the held-out question/text test set for each classifier. For models trained on the KU or OSU subsets, they also report the performance on the respective other full OSU or KU corpus for cross-corpus testing. As before, for each accuracy we also report whether it significantly differs from the majority baseline.

For the held-out question analysis in Table 6, the performance of all three models (KU, OSU, 7K) on the regular test set and on the cross-corpus set is comparable to what we found in Study 1 in Section 6.1. This is to be expected because this part of Study 2 is virtually identical to Study 1 and serves only for comparison with the held-out question test set. For the held-out question test set, we find that performance drops to between the regular test set and the held-out question test

10. The figure again does not show the development set, as in Figure 3.

Table 6. Classifier performance on the held-out question (HoQ) splits of CREG (RTest = regular test set; HoTest = held-out test set)

Trained on	Tested on	Acc.	95% CI	Baseline	P-Value
$KU_{HoQ\text{-}train}$	$KU_{HoQ\text{-}RTest}$	84.56	[82.92; 86.09]	25.89	$< 2^{-16}$
	$KU_{HoQ\text{-}HoTest}$	57.32	[53.59; 61.00]	27.18	$< 2.2^{-16}$
	OSU	27.06	[25.54; 28.62]	27.77	0.8209
$OSU_{HoQ\text{-}train}$	$OSU_{HoQ\text{-}RTest}$	80.57	[77.80; 83.13]	27.73	$< 2^{-16}$
	$OSU_{HoQ\text{-}HoTest}$	61.76	[56.18; 67.11]	28.21	$< 2.2^{-16}$
	KU	28.69	[27.67; 29.72]	25.89	2.056^{-08}
$7K_{HoQ\text{-}train}$	$7K_{HoQ\text{-}RTest}$	74.90	[73.49; 76.26]	26.39	$< 2^{-16}$
	$7K_{HoQ\text{-}HoTest}$	54.43	[50.53; 58.30]	29.66	$< 2.2^{-16}$

set: for KU it drops from 84.56% to 57.32%, for OSU from 80.57% to 61.76%, and for CREG-7K from 74.90% to 54.43%. Throughout, the performance on the held-out question test set is much higher than the majority baseline, confirming some generalization of the model to answers of unseen questions.

For the held-out text classifiers, the results are equivalent, though the performance drop between the regular and the held-out text test set is steeper. Accuracy on KU goes from 86.18% to 40.47%, on OSU from 83.23% to 40.74%, and on CREG-7K from 77.52% to 40.09%. This means that the classifier generalizes to some extent to unseen questions on unseen texts, although it clearly performs best on unseen answers to seen questions and texts.

Table 7. Classifier performance on the held-out text (HoT) splits of CREG (RTest = regular test set; HoTest = held-out test set)

Trained on	Tested on	Acc.	95% CI	Baseline	P-Value
$KU_{HoT\text{-}train}$	$KU_{HoT\text{-}RTest}$	86.18	[84.33; 87.89]	25.77	$< 2^{-16}$
	$KU_{HoT\text{-}HoTest}$	40.47	[36.97; 44.04]	27.15	1.155^{-15}
	OSU	26.36	[24.85; 27.91]	27.77	0.9660
$OSU_{HoT\text{-}train}$	$OSU_{HoT\text{-}RTest}$	83.23	[80.10; 86.05]	26.53	$< 2^{-16}$
	$OSU_{HoT\text{-}HoTest}$	40.74	[35.56; 46.08]	31.62	0.0002
	KU	28.09	[27.08; 29.12]	25.89	7.49^{-06}
$7K_{HoT\text{-}train}$	$7K_{HoT\text{-}RTest}$	77.52	[76.09; 78.91]	27.36	$< 2^{-16}$
	$7K_{HoT\text{-}HoTest}$	40.09	[36.33; 43.94]	29.35	2.408^{-09}

7. Performance of complexity models on learner language

To address our third research question, targeting the robustness of NLP on learner data and its impact on the linguistic complexity analysis, we contrast results based on the automatic NLP analysis with those using the CREG-104 reference annotation introduced in Section 4.3.1. To explore the validity of the NLP analysis of learner data as well as the impact the analysis has on the complexity measures, we look into differences at three crucial stages of the experiment pipeline illustrated in Figure 5.

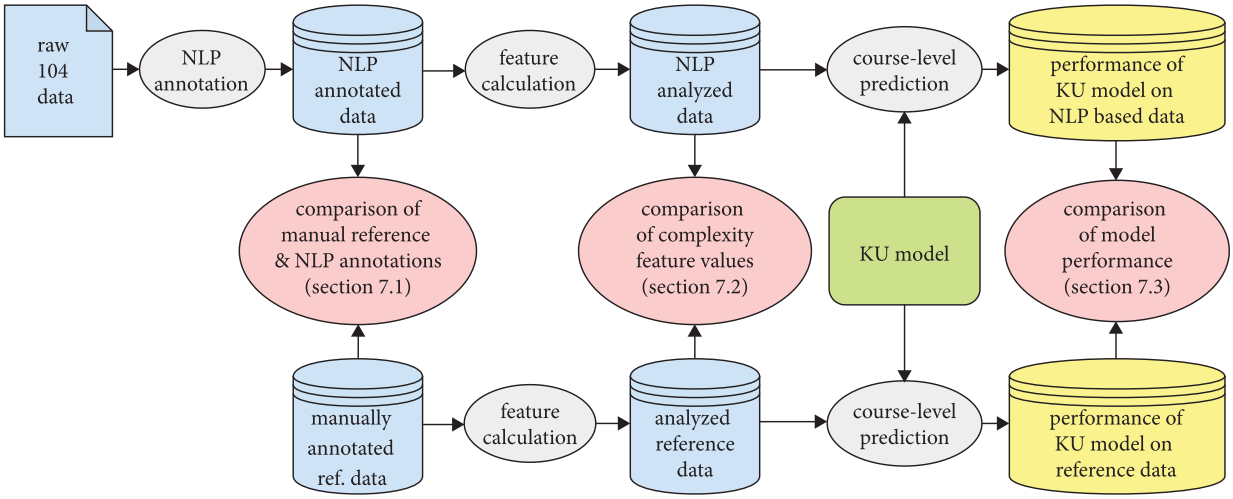


Figure 5. Set-up for RQ3: Comparison of NLP and manual reference annotations (Section 7.1); comparison of feature values extracted based on them (Section 7.2); and model performance using features based on them (Section 7.3)

7.1 Accuracy of NLP analysis

7.1.1 *Set-up of study 3.1*

For the evaluation of the linguistic annotations, we use Mate (cf. Section 5.2) to perform automatic PoS tagging, lemmatization, and morphological analyses on the CREG-104 data using the manual reference tokenization and sentence segmentation. For the evaluation, we calculate the percentage of student answers for which the automatic annotations are fully correct, i.e., the automatic annotation was identical to the manual reference annotation. We also calculate the token-wise accuracy of the automatic annotations for the evaluation of PoS tagging, lemmatization, and morphological analysis. For PoS tagging, we consider the accuracy on the fine-grained STTS tag set (Thielen, Schiller, Teufel, & Stöckert, 1999) containing 56 PoS tags, as well as the accuracy on a more coarse-grained tag set. For the latter we collapsed the classes to distinguish only nouns, verbs, adjectives, adverbs, pronouns,

articles, prepositions, conjunctions, punctuation, particles, foreign material, cardinals, non-words, and separated compound elements. In addition to the overall performance, we also report in more detail the performance for the following lexical PoS categories that play an important role in the calculation of the complexity features: attributive adjectives (adj), adverbs and adjectives with adverbial use (adv), nouns, finite verbs, and non-finite verbs.

For the evaluation of dependencies, we calculate the token-wise Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS), i.e., the percentage of tokens with the correct head and, for LAS, the correct dependency relation label. We again augment the overall evaluation with a more fine-grained analysis looking into the performance on specific dependency labels of particular relevance for the calculation of complexity features in the pipeline: subjects including passivized subjects (SB); clausal, accusative, genitive, and dative objects (OB); relative clauses (RC); conjuncts (CJ); modifiers (MO); and separable verb particles (SVP).

7.1.2 *Results of study 3.1*

Table 8 summarizes the performance of the automatic annotations for PoS, lemmas, and various morphological indices.

Table 8. Accuracy of automatic PoS, morphology, and lemma annotation on CREG-104. Italics indicate where the morphological annotation can be inferred from the PoS tag (DoC = degree of comparison, Verb+F = finite verbs; Verb−F = non-finite verbs)

	Adj. (*N* = 25)	**Adv.** (*N* = 51)	**Noun** (*N* = 239)	**Verb+F** (*N* = 98)	**Verb−F** (*N* = 37)	**Other** (*N* = 449)	**Σ** (*N* = 899)
Parts-of-Speech							
Acc. fine	100	74.51	93.72	94.90	75.68	95.99	93.33
Acc. coarse	100	84.31	98.33	99.00	89.19	96.21	96.22
Lemmas							
Acc. lemma	84.00	96.08	91.63	85.71	89.19	98.66	94.44
Morphology							
Acc. case	80.00	*92.16*	88.28	*100*	*89.19*	96.88	93.88
Acc. number	100	*92.16*	94.56	95.92	*75.68*	98.22	95.77
Acc. gender	96.00	*92.16*	85.36	*100*	*89.19*	95.10	92.66
Acc. person	*100*	*100*	*100*	95.92	*86.49*	100	99.00
Acc. tense	*100*	*100*	*100*	94.90	*86.49*	*100*	98.89
Acc. mode	*100*	*100*	*100*	95.92	*86.49*	*100*	99.00
Acc. DoC	100	86.27	*98.74*	*98.98*	*94.59*	*99.55*	98.33

The overall accuracy across all 899 tokens[11] in CREG-104 in the final column shows that the performance of the Mate tools on the learner data is very good, with accuracy ranging from 92.66% for the morphological analysis of gender to 99.00% for the morphological analysis of person and mode. The PoS tagging of adverbs and non-finite verbs receives the lowest accuracy, whereas the identification of adjectives and finite verbs works particularly well.

For the morphological analysis, unsurprisingly the prediction works best where morphological categories do not apply to a given PoS tag. These cases are shown in italics and contribute most of the 100% accuracy values in the table. For example, German attributive adjectives are inflected according to case, number, gender, and degree of comparison but not person, tense, or mode. Thus, given an adjective PoS tag, the only conceptually possible morphological tag for the latter indices is "not applicable". The accuracy of the morphological analysis for these cases is thus entirely dependent on the performance of the PoS tagging. That this dependence can also have negative effects is illustrated by the non-finite verbs, showing the lowest morphology accuracies. Overall, nine out of the 37 non-finite verbs are incorrectly PoS tagged. Two are tagged as adjectives and two as nouns, receiving corresponding analyses for case, number, and gender. Five are tagged as finite verbs and receive a morphological analysis for number, person, tense, and mode. Looking at the cases where the morphological analyzer cannot solely rely on the PoS tagger, we see that while the identification of number in general and verb inflection on finite verbs performs relatively well, the identification of case and gender on adjectives and nouns seems to be the most challenging to analyze. However, also in these cases the performance of the analysis is reasonably good, never dropping below 80% accuracy. When looking at these results, it should be kept in mind, though, that the analysis is based on a very small reference data set (CREG-104 contains 110 sentences with 780 words). Thus, in terms of morphological categories, only very little of the potential analysis space is in fact represented in CREG-104. This holds specifically for the following morphological indices: (a) person inflection on verbs, because all verbs but one are in third person; (b) degree of comparison on adjectives, because there is only one comparative and one superlative adjective; (c) verb mode, because only three finite verbs are subjunctive, the rest being indicative. However, the data contains a relatively balanced distribution for gender, case, number, and tense features across the appropriate word categories.

Table 9 displays the performance results of the dependency analysis. Overall, the UAS is relatively high, with the exception of relative clauses with a UAS of

11. Standard NLP tokenization separates punctuation from words, so for CREG-104 we obtain 780 words plus 119 punctuation tokens.

75.00%. However, given that there are only four instances of relative clauses in CREG-104, caution is warranted when interpreting this result, as the difference between the observed 75% and 100% is the incorrect attachment of a single relative clause in the data.

Table 9. Performance of automatic dependency annotation on CREG-104 (SB = subject, OB = object, RC = relative clause, CJ = conjunct, MO = modifier, SVP = separable verb particle; UAS = unlabeled attachment score, LAS = labeled attachment score)

	SB (*N* = 98)	**OB** (*N* = 91)	**RC** (*N* = 4)	**CJ** (*N* = 46)	**MO** (*N* = 77)	**SVP** (*N* = 4)	**Rest** (*N* = 579)	**Σ** (*N* = 899)
UAS	94.90	91.21	75.00	82.61	80.52	100	93.44	91.66
LAS	91.84	81.32	50.00	78.26	79.22	25.00	91.54	88.32

Considering the well-established difficulty of labeling dependents, the LAS results are in an acceptable range. Exceptions to this are separable verb particles and the already mentioned relative clause relations. While separable verb particles are more frequently found in the data, it is well-known that their identification is rather unreliable even for standard German data (Weiss, 2015).

While the UAS and LAS are relatively high for subject relations, the LAS for object relations at 81.32% is considerably lower. This should be taken into account when using complexity features that rely heavily on the correct identification of objects, such as the grammatical transition counts among the discourse features.

Finally, it is interesting to compute for which percentage of student answers the entire annotation was correct, i.e., the automatic annotation was identical to the manual reference annotation. For the labeled dependency analysis, this is the case for 53.40% of the student answers. If we also require the PoS analysis to be identical, the percentage drops to 39.84%. Further requiring the lemmas to be identical results in 27.18%. Finally, also requiring the identical morphological annotation for the entire student answer results in 20.36% of the student answers showing exact complete identity of the automatic analysis and the reference annotation across all annotation layers. So while we saw that automatic analysis supports individual high-quality annotation decisions, the overall linguistic annotation of most student answers contains some errors. In the following section, we investigate the impact of these annotation errors on the extraction of complexity features.

7.2 Effect on linguistic complexity analysis

7.2.1 *Set-up of study 3.2*

To investigate the impact of the differences between the automatic and the reference linguistic analysis on the complexity features computed on that basis, we first identified all complexity features that are either fully or partially based on the PoS, lemma, morphology, or dependency annotations ($N = 93$, see Appendix B). We extracted these features from CREG-104, once using the manual reference annotation as the basis for the calculation (henceforth: reference basis) and once using the automatic NLP annotations (henceforth: NLP basis). We then removed all invariable features, i.e., those targeting linguistic phenomena that do not vary on CREG-104. This resulted in a set of 69 complexity features.

For comparing the difference between continuous variables, accuracy is not meaningful, so we calculated the z-scores of the complexity features and then computed the root mean squared difference (RMSD) between the features calculated on the reference basis and the features calculated on the NLP basis. The RMSD measure is inspired by the root mean squared error, the well-established metric used to evaluate regression models predicting continuous variables. The RMSD measure provides a compact summary of the difference in standard deviations between the features calculated based on the two annotation bases (reference vs. NLP), with large differences impacting the measure more than a number of small differences.

7.2.2 *Results of study 3.2*

Table 10 shows the RMSD between the complexity features based on the automatic annotation compared to those based on the CREG-104 reference annotation (see Appendix B for feature definitions).

Since we are comparing the z-scores for each feature, a RMSD of one indicates a difference of one standard deviation between the value of the feature computed on the NLP basis compared to its computation on the reference annotation. Depending on the purpose of an analysis, researchers may find different RMSDs acceptable. In Table 10, no feature shows an extreme[12] difference and only two features have an RMSD of more than 1, both of which are grammatical transition probabilities, from subject to nothing and from object to nothing. The eleven features with a medium RMSD between 0.5 and 1 standard deviations are predominantly based on the assignment of the subject or object dependency relation or associated case

12. We refer to RMSD differences as "extreme" for RMSD > 2, inspired by characterizations of outliers as more than two standard deviations from the mean, "substantial" for RMSD > 1, "medium" > 0.5, "small" ≤ 0.5.

Table 10. RMSD of features calculated on automatic annotations and reference annotations (group abbreviations: DISC = discourse complexity, PHR = phrasal complexity, CLA = clausal complexity, LEX = lexical complexity, MOR = morphological complexity, USE = language use, HP = human processing)

Feature	Group	RMSD
Substantial differences		
Transition probability of object role to none	DISC	1.39
Transition probability of subject role to none	DISC	1.00
Medium differences		
Adjective and adverb verb modifiers per verb phrase	PHR	0.76
Lexical units per synset*	LEX	0.74
Non-subject prefields per prefield†	PHR	0.74
Accusative case per noun	MOR	0.66
Coverage of verb modifier types	PHR	0.61
Sum longest dependency per sentence	HP	0.58
Nominative case per noun	MOR	0.58
Longest dependency	HP	0.57
Dative case per noun	MOR	0.55
Hypernyms per type found in GermaNet	LEX	0.51
Synsets per type found in GermaNet	LEX	0.51
Small differences		
Frames per verb found in GermaNet	LEX	0.50
Relations per sysnset	LEX	0.48
Hyponyms per type found in GermaNet	LEX	0.48
Average number of noun phrase dependents	PHR	0.48
Prepositional verb modifiers per verb	PHR	0.47
Clausal noun modifiers per noun phrase	PHR	0.47
Average number of verb phrase dependents excluding modal verbs	PHR	0.45
Maximal total integration cost per finite verb (original weights)	HP	0.45
Total integration cost per finite verb (original weights)	HP	0.43
Average number of syllables between first argument and verb	PHR	0.42
Maximal total integration cost per finite verb (configuration V)	HP	0.38
Simple present tense per finite verb	MOR	0.37

Table 10. *(continued)*

Feature	Group	RMSD
Coverage of tenses	MOR	0.36
Total integration cost per finite verb (configuration V)	HP	0.36
Maximal total integration cost per finite verb (configuration C)	HP	0.36
Total integration cost per finite verb (configuration C)	HP	0.35
Possessive noun modifiers per noun phrase	PHR	0.35
Verbs with third person marking per finite verb	MOR	0.34
Maximal total integration cost per finite verb (configuration CV)	HP	0.34
Genitive case per noun	MOR	0.33
Coverage of noun modifier types	PHR	0.31
Average number of verb phrase dependents	PHR	0.30
Maximal total integration cost per finite verb (configuration CMV)	HP	0.28
Maximal total integration cost per finite verb (configuration MV)	HP	0.27
Simple past tense per finite verb	MOR	0.27
Maximal total integration cost per finite verb (configuration CM)	HP	0.27
Participle verbs per verb	MOR	0.27
Maximal total integration cost per finite verb (configuration V)	HP	0.26
Log lemma frequency per type found in KCT	USE	0.25
Total integration cost per finite verb (configuration CM)	HP	0.25
Total integration cost per finite verb (configuration CMV)	HP	0.25
Lemma frequency per type found in KCT	USE	0.25
Total integration cost per finite verb (configuration M)	HP	0.25
Total integration cost per finite verb (configuration MV)	HP	0.25
Verbs with indicative marking per finite verb	MOR	0.22
Minimal age of active use for lemma types	USE	0.19
Average age of active use for lemma types	USE	0.19
Lemma frequency per type found in dlexDB	USE	0.18
Percentage of lemmas found in dlexDB	USE	0.14
Percentage of lemmas found in KCT	USE	0.14
sein (to be) instances per verb	LEX	0.13
Transition probability of no role to subject role	DISC	0.11
Log lemma frequency per type found in dlexDB	USE	0.10

Table 10. *(continued)*

Feature	Group	RMSD
Transition probability of no role to object role	DISC	0.08
Derived nouns per noun	MOR	0.07
Maximal age of active use for lemma types	USE	0.02
No difference		
Eventive passive per sentence	PHR	0.00
Global overlap of arguments per sentence	DISC	0.00
Local overlap of arguments per sentence	DISC	0.00
Transition probability of no role to no role	DISC	0.00
Transition probability of no role to other role	DISC	0.00
Transition probability of other role to no role	DISC	0.00
Deverbal nouns per noun	MOR	0.00
Verbs with first person marking per finite verb	MOR	0.00
Verbs with subjunctive marking per finite verb	MOR	0.00
haben (to have) instances per verb	LEX	0.00

* Synsets are sets of semantically-related lexical units in GermaNet.
† The prefield is the area preceding the finite verb in a German main clause, see Wöllstein (2014).

markings such as accusative and dative – for which we found lower NLP accuracies in Section 7.1. In addition, we also find features in this medium group that are based on GermaNet synset measures and rely on lemmatization, verb modification, and maximal dependency length.

Most of the features (56 out of 69) show only a weak difference (46/56) or no difference at all (10/56) between the calculation on both feature bases. As expected, this predominantly includes features based on linguistic annotations that the automatic analysis can provide with high accuracy (cf. Section 7.1), such as lemma-based frequency and auxiliary verb ratios, DLT features and comparable argument-verb distance and overlap measures as well as verb and noun phrase dependency features relying predominantly on correct dependency attachments, and morphological measures of different types of verb inflections. Interestingly, the set also includes some measures relying on linguistic annotations that we had found performing relatively poorly, e.g., the clausal noun phrase modifier feature based on the relative clause label or features based on genitive case markings on nouns.

Whether the quality of complexity features computed on the basis of automatic NLP analysis is sufficient depends on the purpose for which they are to be used. In this article, we use complexity features to predict the course levels of student writings. To identify whether the automatically computed complexity features are

acceptable for this purpose or not, the next section reports on the third part of Study 3 comparing classifier performance on test sets using automatic annotation and manual reference annotation.

7.3 Effect on proficiency classification

7.3.1 *Set-up of study 3.3*

For the final machine learning experiment, we used the CREG-KU data with the automatically computed complexity analysis. We reduced the feature set to the 69 features that are based exclusively or partially on PoS, lemma, morphology, and dependency annotations and were variable on CREG-104. We calculated the z-scores of the features and split the CREG-KU data into training, development, and test sets using a 70/20/10 split. On this basis, we again trained an ordinal random forest classifier and evaluated it on three test sets: (a) the regular test set with fully automatically extracted complexity features; (b) the CREG-104 data with fully automatically extracted complexity features; and (c) the CREG-104 data with complexity features extracted on the reference annotation basis.[13]

7.3.2 *Results of study 3.3*

Table 11 shows the overall course-level prediction accuracy of the classifier on the three test sets and the majority baselines. For CREG-104, the baseline is higher since it is not balanced for course levels. The classifier trained on CREG-KU significantly outperforms the majority baseline on all test sets. For the CREG-104 test set, the performance is slightly higher for the NLP-based features (78.85) than for the reference-based ones (73.08), possibly because the training set features are also NLP-based, but the overlapping confidence intervals ([69.74;86.24] and [63.49;81.31]) show that this is not a reliable difference we can meaningfully investigate based on this data set.

Table 11. Classification performance of a 69-feature model trained on CREG-KU when applied to complexity features based on automatic and manual reference annotations

Tested on	Feature basis	Acc.	95% CI	Baseline	P-Value
$\text{CREG-KU}_{\text{test}}$	NLP	74.50	[72.45; 76.47]	28.19	$< 2^{-16}$
CREG-104	NLP	78.85	[69.74; 86.24]	45.19	2.291^{-12}
CREG-104	reference	73.08	[63.49; 81.31]	45.19	$7{:}532^{-09}$

13. The answers in CREG-104 were also written at KU but are not part of this CREG-KU data set.

Table 12 shows the classifier's performance on the three test sets split by course level. Since we now turn to a discussion of the classification performance at individual course levels, we report precision, recall, and F1 score of the course-level prediction instead of accuracy (for definitions, see Tharwat, 2018). On the CREG-KU_{test} test data set we observe a comparable precision across course levels, but a higher recall for the two A1 level courses, leading to an overall better performance in terms of F1 scores on course levels A1.1 and A1.2. This does not carry over to CREG-104, though, where the performance is generally higher for course level A2.2 for both feature bases. This is likely to be an artifact of the data imbalance in CREG-104, which contains about 45% A2.2 texts.

Since CREG-104 does not contain any A1.2 course-level data, we cannot calculate recall and F1 score there, and the few answers incorrectly labeled as A1.2 naturally result in a precision of 0%. We thus focus the discussion on the other three course levels. When comparing differences in the course-level-wise performance for the NLP-based features on CREG-KU_{test}, we find that the performance does not specifically suffer for lower levels, which would have indicated specific NLP problems with beginning learners and their creative forms. At the same time, the proficiency range in the data is limited here, so we cannot fully investigate how the NLP quality develops as the proficiency increases towards fully well-formed but also more complex language, where the former makes the NLP easier, but the latter makes it harder.

When comparing the NLP-based with the reference-based features on CREG-104, we find that for course level A1.1 and A2.2, there is virtually no difference in precision between the two. However, the recall is much higher with NLP-based features. For course level A2.1, both precision and recall are higher when using NLP-based features, too. This indicates that the systematic nature of the automatic annotation on the training and test data is more important for the classifier than having access to higher quality linguistic analysis on the test data only. A comparison with the performance when training and testing on manually annotated reference data unfortunately is not feasible given the cost of producing a sufficiently large manually annotated corpus.

Overall, the generally comparable performance across proficiency levels for using manual or automatic annotation methods supports the assumption that NLP-based complexity analyses can provide valuable evidence for building classifiers capable of identifying proficiency differences and development.

Table 12. Precision, recall, and F1 score for course-level prediction using NLP-based and reference-based features

Tested on	Feature basis	Measure	A1.1	A1.2	A2.1	A2.2
CREG-KU$_{test}$	NLP	Precision	75.44	74.51	74.63	73.07
		Recall	78.66	79.92	67.18	70.44
		F1 score	77.01	77.12	70.71	71.73
CREG-104	NLP	Precision	76.92	0.00	82.76	95.00
		Recall	80.00	n.a.	75.00	80.85
		F1 score	78.43	n.a.	78.69	87.36
	Reference	Precision	76.00	0.00	78.57	94.59
		Recall	76.00	n.a.	68.75	74.47
		F1 score	76.00	n.a.	73.33	83.33

8. Discussion

The classification experiments in the first two studies show that the language evidence contained in the short answers to reading comprehension exercises supports high-quality L2 proficiency classification using a broad range of linguistic complexity measures. Our first research question thus can be answered affirmatively. About half of the linguistic complexity features (marked by superscript a in Appendix B) are informative for such short answers, and training machine learning models on that basis results in classifiers with accuracies of over 81% for each school, and over 73% for the combined data – a clear success in relation to majority baselines of under 28%. Complexity analysis and classification can therefore be meaningfully applied to short answer tasks, which extends the reach and empirical basis of linguistic complexity research.

Regarding the second research question, the extent to which features and models generalize across task contexts as made explicit by different questions and reading texts, the results are varied. While we obtain high accuracies for answers to questions and texts for which other answers are included in the training data, we see a clear drop in accuracy when testing on answers to unseen questions and even more so for unseen texts. Interestingly, a model can be trained that successfully performs for the data of both universities together, which confirms the generalizability of the feature set. But the models trained on the data from one university hardly generalize at all to that of the other university. The data from the two universities thus seem to cover different parts of the distribution of linguistic complexity characteristics.

To better understand the nature of this distribution, in the future we need to investigate the nature of the relevant task context characteristics and their effect on the linguistic complexity of the learner answers they elicit. For example, the KU and OSU data sets may differ in the question types they contain, both in terms of the question form and in terms of the operators they express – and some of those tasks may allow or require learners to use more complex language than others. So as it stands, our analysis is limited in not differentiating between complexity arising from increasing proficiency and that arising from tasks supporting a more complex use of language. Future work should try to address to what extent the discrimination of course levels can be maintained when the same reading exercises are administered across course levels – though it naturally will be difficult to compile a corpus using tasks that can meaningfully be administered across a range of proficiency levels. Similarly, the variability of task contexts researched here could be extended to consider the effect of different text genres in future work (e.g., descriptive vs. expository texts).

For the third question asking about the impact of the characteristics of learner language on linguistic complexity classification, we looked at three stages: the linguistic analysis, the calculation of complexity features on that basis, and the statistical modeling based on these features. We found that for the linguistic properties of relevance for our linguistic complexity measures, the performance of the NLP methods in general is quite close to the manual linguistic analysis. Beyond lending some legitimacy to the use of automated complexity analysis, the results are substantially more fine-grained. For example, in the typical NLP pipeline architectures, the dependency between the linguistic analysis levels needs to be kept in mind.[14] The morphological analyzer depends on the PoS tagger, and the dependency parser relies on the morphology and PoS analysis. As a result, even partial manual reference annotation for the lower levels of the NLP pipeline, such as PoS tagging, can have a substantial impact on the overall performance on learner data.

In a similar vein, the more NLP steps required, the more error prone the analysis will be. In our analysis, the identification of dependency labels was one of the most error prone analyses. When considering how to increase the reliability of complexity feature computation, we therefore should avoid reference to dependency relations and instead express them in terms of more basic, reliable annotation where possible – which is reminiscent of the process of translating research questions into the annotations available in a given corpus (Meurers, 2005). For example, instead of making reference to the relative clause dependency label, which we found to be

14. The CoALLa project (http://purl.org/coalla) explores an alternative architecture for analyzing learner language that more readily supports integration of bottom-up form-based and top-down task/meaning-based information.

assigned in learner language with limited accuracy, the complexity code refers to subordinate clauses introduced by relative pronouns identified via PoS tagging. For features that cannot be calculated without elaborate NLP analyses, caution is warranted when interpreting them for learner data, such as the grammatical transition counts including object roles in our study.

9. Conclusion

We investigated the applicability of complexity measures to short answers to reading comprehension questions produced by L2 learners of German and the robustness of automatic complexity modeling on learner data. We showed that the limited linguistic evidence provided by this type of short written learner production is sufficient to build successful predictive models of L2 proficiency. Capturing linguistic complexity with a broad range of features across all domains of the linguistic system and language use thus pays off in capturing sufficient evidence even when limited language data is available. At the same time, the weighting of the evidence in the trained models is substantially dependent on the characteristics of the task context. We found that the models still generalize somewhat to unseen questions on seen reading texts, but much less when applied to data from completely new reading tasks that have nothing in common with the tasks that resulted in the training data. In line with linguistic complexity research on task effects (Alexopoulou et al., 2017; Biber et al., 2016; Caines & Buttery, 2017; Michel et al., 2019), we view this as evidence for the need to interpret linguistic complexity in relation to the properties of the task that elicited the data.

In terms of analysis methodology, automating the linguistic analysis worked remarkably well for beginner German learner data. The effect of using either the manual reference or the automatic annotations for the extraction of complexity features was weak, with only a small impact on the final proficiency classification model. However, we also found that errors in the early stages of the NLP pipeline percolate up to later analyses, making it advisable to base linguistic complexity analysis on lower levels of NLP analysis where this is possible.

Funding

The work of the project A4 (http://purl.org/comic) was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — SFB 833 – Project ID 75650358.

Acknowledgements

We are grateful to Ramon Ziai, Niels Ott, and Björn Rudzewitz for the task-based CREG corpus collection and the creation of the WELCOME tool facilitating this in an authentic language teaching context. We would particularly like to thank Kathy Corl and her group at The Ohio State University and Nina Vyatkina and her team at the University of Kansas for the collection and assessment of the learner data.

References

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67, 181–209. https://doi.org/10.1111/lang.12232

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5), 639–668. https://doi.org/10.1093/applin/amu059

Björkelund, A., Bohnet, B., Hafdell, L., & Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Demonstration volume of the 23rd COLING* (pp. 23–27). Beijing.

Bohnet, B., & Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 joint conference on EMNLP and computational natural language learning* (pp. 1455–1465). Jeju Island, Korea: Association for Computational Linguistics.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The TIGER treebank. In *Proceedings of the workshop on treebanks and linguistic theories*. Sozopol.

Brants, T., Skut, W., & Uszkoreit, H. (1999). Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA treebank workshop*. Paris.

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written l2 texts. *Second Language Research*, 35(1), 99–119. https://doi.org/10.1177/0267658316643125

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2), 540–545. https://doi.org/10.3758/BRM.40.2.540

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58, 412–424. https://doi.org/10.1027/1618-3169/a000123

Caines, A., & Buttery, P. (2017). The effect of task and topic on opportunity of use in learner corpora. In *Learner corpus research: New perspectives and applications*. London: Bloomsbury.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on EMNLP* (pp. 740–750). Doha, Qatar.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S.A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading*, 42(3–4), 541–561. https://doi.org/10.1111/1467-9817.12283

Crossley, S.A., Weston, J.L., Sullivan, S.T.M., & McNamara, D.S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3), 282–311. https://doi.org/10.1177/0741088311410188

De Clercq, B., & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research Special Issue on Linguistic Complexity*, 35(1), 71–97.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2014). Assessing document and sentence readability in less resourced languages and across textual genres. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of the International Journal of Applied Linguistics*, 165(2), 163–193.

Díaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2), 139–154.

Duden. (2009). *Deutsche Grammatik* (4th ed., Vol. 4). Dudenverlag.

Ellis, N.C. (2002). Frequency effecs in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, UK: Oxford University Press.

François, T., & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 joint conference on EMNLP and computational natural language learning*.

Galasso, S. (2014). Exploring textual cohesion characteristics for German readability classification (Bachelor Thesis in Computational Linguistics). Department of Linguistics, University of Tübingen. (http://purl.org/dm/papers/Galasso-14.pdf)

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st SLRF*. Cascadilla Press.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: papers from the first mind articulation project symposium* (pp. 95–126). MIT.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. *Proceedings of the 8th International Language Ressources and Evaluation*, 759–765.

Hamp, B., & Feldweg, H. (1997). GermaNet – a lexical-semantic net for German. In *Proceedings of ACL workshop automatic information extraction and building of lexical semantic resources for NLP applications*. Madrid.

Hancke, J. (2013). Automatic prediction of CEFR proficiency levels based on linguistic features of learner language (Unpublished master's thesis). Department of Linguistics, University of Tübingen.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th COLING* (pp. 1063–1080). Mumbay, India.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, 62, 10–20. https://doi.org/10.1026/0033-3042/a000029

Höhle, T. N. (1986). Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne (Ed.), *Kontroversen alte und neue. Akten des VII. Internationalen Germanistenkongresses Göttingen 1985* (pp. 329–340). Tübingen: Niemeyer. (Bd. 3)

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research. Special Issue on Linguistic Complexity*, 35(1), 2–31.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 1–20). John Benjamins. https://doi.org/10.1075/lllt.32.01hou

Hunt, K. W. (1965). A synopsis of clause-to-sentence length factors. *The English Journal*, 54(4), 300+305-309. https://doi.org/10.2307/811114

Lavalley, R., Berkling, K., & Stüker, S. (2015). Preparing children's writing database for automated processing. In *Proceedings of the workshop on language teaching, learning and technology at speech and language technologies in education* (pp. 9–15).

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In M. Walter & P. Grommes (Eds.), *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweispracherwerbsforschung* (pp. 119–140). Tübingen: Max Niemeyer Verlag.

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. In *Proceedings of corpus linguistics*. Birmingham.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD) (Unpublished doctoral dissertation). University of Memphis.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Meurers, D. (2005). On the use of electronic corpora for theoretical linguistics. case studies from the syntax of German. *Lingua*, 115(11), 1619–1639. https://doi.org/10.1016/j.lingua.2004.07.007

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The cambridge handbook of learner corpus research* (pp. 537–566). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.024

Meurers, D. (2020). Natural language processing and language learning. In C. A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 817–831). Oxford: Wiley.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(2). https://doi.org/10.1111/lang.12233

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, 3, 124–152. https://doi.org/10.1558/isla.38248

Ott, N., & Ziai, R. (2010). Evaluating dependency parsing performance on German learner language. In M. Dickinson, K. Müürisep, & M. Passarotti (Eds.), *Proceedings of the ninth international workshop on treebanks and linguistic theories* (Vol. 9, pp. 175–186). Tartu, Estonia: Tartu University Press. http://hdl.handle.net/10062/15960

Ott, N., Ziai, R., & Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 47–69). Amsterdam: Benjamins. https://doi.org/10.1075/hsm.14.05ott

Petrov, S., & Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of the NAACL main conference* (pp. 404–411). Rochester, New York.

Pilán, I., Vajjala, S., & Volodina, E. (2015). A readable read: Automatic assessment of language learning materials based on linguistic complexity. In *Proceedings of CICLING 2015*.

Reis, M. (2001). Bilden Modalverben im Deutschen eine syntaktische Klasse? In R. Müller & M. Reis (Eds.), *Modalität und Modalverben im Deutschen*. Hamburg: Helmut Buske. (Linguistische Berichte – Sonderhefte)

Seeker, W., & Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a German treebank. In *Proceedings of the 8th international conference on language resources and evaluation* (pp. 3132–3139). Istanbul, Turkey.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the workshop on computational linguistics for linguistic complexity* (p. 49–58). Osaka.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, 33(2), 149–183. https://doi.org/10.1177/0741088316631527

Tagliamonte, S.A. (2011). *Variationist sociolinguistics: Change, observation, interpretation*. John Wiley & Sons.

Telljohann, H., Hinrichs, E., & Kübler, S. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the fourth LREC*. Lissabon.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*.

Thielen, C., Schiller, A., Teufel, S., & Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS* (Tech. Rep.). Stuttgart/Tübingen: Institut für Maschinelle Sprachverarbeitung Stuttgart and Seminar für Sprachwissenschaft Tübingen.

Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1), 58–95. https://doi.org/10.1075/ijlcr.1.1.03tra

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh BEA workshop* (pp. 163–173).

Weiss, Z. (2015). More linguistically motivated features of language complexity in readability classification of German textbooks: Implementation and evaluation (Bachelor's Thesis). Department of Linguistics, University of Tübingen. (http://purl.org/zweiss/rsrc/Weiss-15-BA-CL.pdf)

Weiss, Z. (2017). Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects (Unpublished master's thesis). University of Tübingen, Germany. (http://purl.org/zweiss/ma-thesis/weiss2017-distr.pdf)

Weiss, Z., & Meurers, D. (2018). Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th COLING*. Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026

Weiss, Z., & Meurers, D. (2019a). Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the 14th BEA workshop*. Florence, Italy. https://doi.org/10.18653/v1/W19-4440

Weiss, Z., & Meurers, D. (2019b). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the scope of learner corpus research. Selected papers from the fourth learner corpus research conference*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Wöllstein, A. (2014). *Topologisches Satzmodell* (2nd ed.). Heidelberg: Winter.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66, 130–141. https://doi.org/10.1016/j.system.2017.03.007

Yoon, H.-J., & Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, 275–301.

Ziai, R. (2018). Short answer assessment in context: The role of information structure (Unpublished doctoral dissertation). Eberhard-Karls Universität Tübingen.

Ziegler, N. (2018). Pre-task planning in L2 text-chat: Examining learners' process and performance. *Language Learning & Technology*, 22(3), 193–213.

A. Definition of linguistic units used for the calculation of complexity features

Clauses

All maximal projections of finite verbs and elliptical constructions with sentential status (= all sub-trees tagged with "S"), and to infinitives with sentential status.

Complex t-units

T-units that include subordinate clauses.

Conjunctional clauses

Dependent clauses introduced by a subordinating conjunction.

Dependent clauses with/without conjunction

Conjunctional, interrogative, and relative clauses. Dependent clauses without conjunction are mostly dependent main clauses.

(Graphematic) sentences

Strings containing at least one token that are ended by sentence ending punctuation marks ("!", ".", and "?") or the end of a text.

Half modals

The verbs *haben* (to have), *sein* (to be), *scheinen* (to seem), *drohen* (to threaten), and *versprechen* (to promise), if they govern an infinitive with *zu* (§ 101 Duden, 2009: 101), e.g., *ist zu machen* (is to be done), *droht zu schneien* (threatens to snow). For details, see Weiss (2015: 32).

Lexical words

Nouns, adjectives, adverbs, foreign words, numbers, main verbs, and modal verbs. There is an ongoing discussion on whether modals actually qualify as lexical words (Reis, 2001). Therefore, features using lexical words are partially calculated twice, with and without modal verbs.

Quasi passives

bekommen (to get), *erhalten* (to receive), or *kriegen* (to get) if they govern a past participle (§ 179 Duden, 2009: 147), e.g., *kriegt erklärt* (to get sth. explained).

T-units

"[O]ne main clause plus whatever subordinate clauses happen to be attached to or embedded within it." (Hunt, 1965: 305).

B. List of complexity features used

Table 13. Complexity feature definitions and a reference providing additional information on their implementation, where in column Ref. 1 = Hancke (2013), 2 = Weiss (2015), 3 = Galasso (2014), 4 = Weiss (2017). Features available in different configurations are collapsed into a single item and an example configuration is provided; configuration options and additional specifications are in italics. Superscript *a* indicates use of feature in Studies 1 and 2 in Section 6, *b* indicates that a feature is dependent on the NLP pipeline components evaluated in Study 3 in Section 7

Feature name	Calculation	Ref.
Clausal complexity (25)		
Sum of non-terminal (NT) nodes per sentence [a]	Sum of NT nodes across all constituency parses / number of sentences	1
Sum of non-terminal (NT) nodes per word [a]	Sum of NT nodes across all constituency parses / number of words	1
Average parse tree height per sentence [a]	Summed height of all constituency parses / number of sentences	1
Words per sentence [a]	Number of words / number of sentences	1
Coordinated phrases per sentence [a]	Number of coordinated phrases / number of sentences	1
Complex t-units per sentence	Number of complex t-units / number of sentences	1
T-units per sentence [a]	Number of t-units / number of sentences	1
Clauses per sentence [a]	Number of clauses / number of sentences	1

Table 13. *(continued)*

Feature name	Calculation	Ref.
Dependent clauses per sentence	Number of dependent clauses / number of sentences	1
Conjunctional clauses per sentence	Number of clauses with conjunction / number of sentences	1
Dependent clauses with conjunction per sentence	Number of dependent clauses with conjunction / number of sentences	1
Dependent clauses without conjunction per sentence	Number of dependent clauses without conjunction / number of sentences	1
Interrogative clauses per sentence	Number of interrogative clauses / number of sentences	1
Relative clauses per sentence	Number of relative clauses / number of sentences	1
Sentential to infinitives per sentence	Number of infinitives with the status of a sentence / number of sentences	1
Longest sentence in words [a]	Maximal observed number of words in a sentence	1
Verbs per sentence [a]	Number of verbs / number of sentences	1
Verbs per t-unit [a]	Number of verbs / number of t-units	1
Verbs per clause [a]	Number of verbs / number of clauses	1
Verbs per finite clause [a]	Number of verbs / number of finite clauses	1
Noun phrases (NPs) per sentence [a]	Number of NPs / number of sentences	1
Prepositional phrases (PPs) per sentence [a]	Number of PPs / number of sentences	1
Verb phrases (VPs) per sentence [a]	Number of VPs / number of sentences	1
To infinitives per sentence	Number of to infinitives / number of sentences	1
Complex noun phrases (NPs) per sentence [a]	Number of complex NPs / number of sentences	1
Phrasal complexity (47)		
Average noun phrase length [a]	Sum of words in noun phrases / number of noun phrases	1
Average verb phrase length [a]	Sum of words in verb phrases / number of verb phrases	1
Average prepositional phrase length [a]	Sum of words in prepositional phrases / number of verb phrases	1
Average number of noun phrase dependents [a, b]	Sum of noun phrase dependents / number of noun phrases with dependents	1
Average number of verb phrase dependents [a, b]	Sum of verb phrase dependents / number of verb phrases with dependents	1
Average number of verb phrase dependents excluding dependents excluding modal verbs [a, b]	Sum of verb phrase modal verbs / number of verb phrases with dependents excluding modal verbs	1
Average number of noun phrase modifiers [a]	Sum of noun phrase modifiers / number of noun phrases	2
Average number of verb phrase modifiers	Sum of verb phrase modifiers / number of verb phrases	2
Adjective and adverb verb modifiers per verb phrase [a, b]	Number of adjective and adverb verb modifiers / number of verb phrases	2

Table 13. *(continued)*

Feature name	Calculation	Ref.
Participle verb modifiers per verb [a, b]	Number of participle verbs / number of verbs	2
Prepositional verb modifiers per verb [a, b]	Sum of prepositional verb modifiers / number of verbs	2
Average verb cluster size	Sum of verbs forming a verb cluster / number of verbs	2
Number of main verb clusters per verb cluster	Sum main verbs governing another verb / number of verb cluster	2
Number of auxiliary verb clusters per verb cluster	Sum auxiliary verbs governing another verb / number of verb cluster	2
Number of modal verb clusters per verb cluster	Sum modal verbs governing another verb / number of verb cluster	2
Coverage of verb modifier types [a, b]	Number of verb modifier types observed at least once / number of verb modifier types distinguished in the code (= 5; *adverbial verb modifiers, adjectival verb modifiers, prepositional verb modifiers, participle 1 modifiers, participle 2 modifiers*)	2
Standard deviation of verb cluster sizes	Standard deviation of verb cluster sizes *(cut at verb cluster size > 6)*	2
Coverage of verb cluster sizes [a]	Number of verb cluster sizes observed at least once / number of verb cluster sizes distinguished in the code (= 5; *minimal verb cluster with 2 verbs, verb cluster with 3 verbs, verb cluster with 4 verbs, verb cluster with 5 verbs, verb cluster with six or more verbs*)	2
Coverage of verb cluster types [a]	Number of verb cluster types observed at least once / number of verb cluster types distinguished in the code (= 3; *auxiliary verb cluster, modal verb cluster, main verb cluster)*	2
Number of attributive participles per noun phrase	Sum attributive participles modifying nouns / number of noun phrases	2
Clausal noun modifiers per noun phrase [a, b]	Number of clausal noun modifiers / number of noun phrases	2
Comparative noun modifiers per noun phrase	Number of comparative noun modifiers / number of noun phrases	2
Determiners per noun phrase [a]	Number of determiners / number of noun phrases	2
Possessive noun modifiers per noun phrase [a, b]	Number of possessive noun modifiers / number of noun phrases	2
Prenominal modifiers per noun phrase [a]	Number of prenominal noun modifiers / number of noun phrases	2
Postnominal modifiers per noun phrase [a]	Number of postnominal noun modifiers / number of noun phrases	2

Table 13. *(continued)*

Feature name	Calculation	Ref.
Coverage of noun modifier types *a, b*	Number of noun modifier types observed at least once / number of noun modifier types distinguished in the code (= 9; *determiner, possessive noun modifiers, pronominal noun modifiers, postnominal noun modifiers, attributive participle 1 modifiers, attributive participle 2 modifiers, appositions or parentheses, comparative noun modifiers, clausal noun modifiers*)	2
Eventive passive per sentence *b*	Number of eventive passive constructions / number of sentences	1
Periphrastic tenses per finite verb	Number of periphrastic tense constructions / number of finite verbs	2
Tense X per finite verb *a, b for X∈{s.pres,s.past}*	Number of tense X constructions / number of finite verbs; tense construction options for X: *Präsens (simple present), Präteritum (simple past), Perfekt (present perfect), Plusquamperfekt (past perfect), Futur 1 (future 1), Futur 2 (future 2)* Ex.: simple present tense per finite verb	2
Coverage of tenses *a, b*	Number of tenses observed at least once / number of tenses in German (= 6: *simple present, simple past, present perfect, past perfect, future 1 and future 2*)	2
Coverage of periphrastic tenses	Number of periphrastic tenses observed at least once / number of periphrastic tenses in German (= 4: *present perfect, past perfect, future 1 and future 2*)	2
Average middle field length in syllables *a*	Sum of syllables across all middle fields / number of middle fields (Höhle, 1986)	2
Average number of syllables between first argument and verb *a, b*	Sum of syllables between the first argument of a verb and the verb if the first argument does not immediately precede the verb / number of verbs not immediately preceded by their first argument	2
Non-subject prefields per prefield *b*	Number of prefields not occupied by the subject of the clause / number of prefield	2
man (someone) occurrences per subject	Number of *man* (someone) occurrences / number of subjects	2
Infinitival constructions per verb phrase (VP)	Number of infinitive verbs / number of VPs	
lassen occurrences per ver phrase (VP)	Number of *sich- lassen* constructions / number of VPs	2
Half modal clusters per verb phrase (VP)	Number of half modals / number of VPs	2
Passives per sentence	Number of passives / number of sentences	2
Quasi-passives per sentence	Number of quasi passives / number of sentences	2

Table 13. *(continued)*

Feature name	Calculation	Ref.
Coverage of deagentivation patterns [a]	Number of deagentivation patterns observed at least once / number of deagentivation patterns distinguished in the code (= 10: *man occurrences, sich-lassen occurrences, infinitival constructions, half modal clusters, passives, quasi passives, participle-I modifiers, participle-II modifiers, attributive participle-I modifiers, attributive participle-II modifiers*)	2
Lexical complexity (34)		
Syllables per word [a]	Number of syllables / number of words	
Characters per word [a]	Number of characters / number of words	1
Longest word in syllables [a]	Maximal number of syllables in a word	1
Type-token ratio (TTR) [a]	Number of types / number of tokens	1
Root TTR [a]	$\sqrt{TTR}$	1
Corrected TTR [a]	$\sqrt{\frac{Number\ of\ types}{2^{*}number\ of\ tokens}}$	1
Bilogarithmic TTR [a]	log(number of types) / log(number of tokens)	1
Uber index [a]	log(number of tokens)^2 / log(TTR)	1
Yule's K [a]	$10^4 \frac{\Sigma(fX^{*}X^2) - N}{N^2}$, for N = number of tokens, X = vector of frequencies for each type, and fX = frequency of each type frequency in X (see https://cran.r-project.org/web/packages/koRpus)	1
HD-D	see McCarthy and Jarvis (2010)	1
MTLD [a]	see McCarthy and Jarvis (2010)	1
Lexical types per lexical tokens	Number of lexical types / number of lexical tokens	1
Lexical types per token [a]	Number of lexical types / number of tokens	1
Lexical verb types per lexical tokens (including modals) [a]	Number of main and modal verb types / number of main and modal tokens	1
Lexical verb types per lexical verb tokens (including modals) [a]	Number of main and modal verb types / number of main and modal verb tokens	1
Squared lexical verb types per lexical verb tokens (including modals) [a]	(Number of main and modal verb types / number of main and modal verb tokens)^2	1
Corrected lexical verb types per lexical verb tokens (including modals) [a]	Number of lexical verb types / $\sqrt{2^{*}\ number\ of\ lexical\ verb\ tokens}$	
Lexical types per lexical tokens (including modals) [a]	Number of lexical types / number of lexical tokens (including modals)	1
sein (to be) instances per verb [a, b]	Number of *sein* (to be) instances / number of verbs	1
haben (to have) instances per verb [b]	Number of *haben* (to have) instances / number of verbs	1

Table 13. *(continued)*

Feature name	Calculation	Ref.
Nouns per lexical tokens [a]	Number of nouns / number of lexical tokens	1
Nouns per token [a]	Number of nouns / number of tokens	1
Verbs per noun [a]	Number of verbs / number of nouns	1
Adjectives per lexical token [a]	Number of adjectives / number of lexical tokens	1
Adverbs per lexical token [a]	Number of adverbs / number of lexical tokens	1
Adjectives and adverbs per lexical token [a]	Number of adjectives and adverbs / number of lexical tokens	1
Modal verbs per verb	Number of modal verbs / number of verbs	1
Auxiliary verbs per verb [a]	Number of auxiliary verbs / number of verbs	1
Hypernyms per type found in GermaNet [a, b]	Sum of hypernyms of all lemma types in the text in GermaNet / number of lemma types found in GermaNet	1
Hyponyms per lemma type found in GermaNet [a, b]	Sum of hyponyms of all lemma types in the text in GermaNet / number of lemma types found in GermaNet	1
Synset per type found in GermaNet [a, b]	Sum of synsets (= word senses) of all lemma types in the text in GermaNet / number of lemma types found in GermaNet	1
Lexical units per synset [a, b]	Sum of number of lexical units per synyset (=word senses) / number of synsets retrieved for lemma types in text from GermaNet	1
Relations per synset [a, b]	Sum of number of lexical relations for each retrieved synset (=word sense) / number of synsets retrieved for lemma types in text from GermaNet	1
Frames per verb found in GermaNet [a, b]	Sum of number of verb frames (= subcategorization information of a verb) found for all verb lemma types / number of verb lemma types found in GermaNet	1
Morphological complexity (41)		
Nominalizations using suffix X per word	Number of nouns ending in suffix X / number of word tokens; suffix options for X: *-ei, -ling, -heit, -keit, -nis, -ung, -werk, -wesen, -schaft, -tum, -ant, -atur, -ator, -arium, -at, -eur, -ent, -enz, -ast, -ist, -ität, -ismus, -ion, -ur* (including all their inflected forms) Ex.: *-keit* nominalizations per word	1
Average compound depth [a]	Sum of number of compound elements in nouns with more than one compound element / number of compound nouns	1
Compound nouns per noun [a]	Number of compound nouns / number of nouns	1
Derived nouns per noun [a, b]	Number of derived nouns / number of nouns	1
Deverbal nouns per noun [b]	Number of nouns derived from verbs / number of nouns	1
X case per noun [b for all X, a for X∈{nom.,acc.,dat}]	Number of nouns with case X / number of nouns; case options for X: *nominative, accusative, dative, genitive* Ex.: accusative case per noun	1
Finite verbs per verb [a]	Number of finite verbs / number of verbs	1

Table 13. *(continued)*

Feature name	Calculation	Ref.
Non-finite verbs per verb	Number of non-finite verbs / number of verbs	1
Participle verbs per verb [b]	Number of participle verbs / number of verbs	1
Verbs with verb mode X per number of finite verbs [a for X= indicative, b for all X]	Number of verbs with markings for verb mode X / number of finite verbs; verb mode options for X: *imperative, subjunctive, indicative* Ex.: verbs with subjunctive marking per finite verb	1
Verbs with X person marking per number of finite verbs [a for X=3rd, b for all X]	Number of verbs with person marking X / number of finite verbs; person marking options for X: *1st, 2nd, 3rd* Ex.: verbs with first person marking per finite verb	1
Discourse complexity (64)		
Local overlap of linguistic material X per sentence [b for X∈{arg.,content,stem}]	Number of instances where linguistic material is repeated in two adjacent sentences / number of sentences; linguistic material options for X: *nouns, arguments, content words, content word stems* Ex.: Local overlap of arguments per sentence	3
Global overlap of linguistic material X per sentence [b for X∈{arg.,content,stem}]	Number of instances where linguistic material is repeated across any sentence in the text / number of sentences; linguistic material options for X: see above Ex.: Global overlap of arguments per sentence	3
Transition probability of grammatical role A to grammatical role B [b for all X]	Number of transitions between adjacent sentences of entities from grammatical role A to grammatical role be in the entire text / (number of sentences – 1) * number of entities in the text; grammatical roles considered: *subject, object, other complement, not present in the sentence* Ex.: probability of transition from subject role to object role	3
Pronouns per noun [a]	Number of pronouns / number of nouns	3
Personal pronouns per noun [a]	Number of personal pronouns / number of nouns	3
Possessive pronouns per noun	Number of possessive pronouns / number of nouns	3
Personal pronouns with person X per noun [a for X∈{3rd}]	Number of personal pronouns with person X / number of nouns; person options for X: *1st, 2nd, 3rd* Ex.: third person personal pronouns per noun	3
Possessive pronouns with person X per noun	Number of possessive pronouns with person X / number of nouns; person options for X: see above Ex.: third person possessive pronouns per noun	3
Personal or possessive pronouns with person X per noun [a for X∈{3rd}]	Number of personal or possessive pronouns with person X / number of nouns; person options for X: see above Ex.: third person personal or possessive pronouns per noun	3
Definite articles per article [a]	Number of definite articles / number of articles	3
Indefinite articles per article	Number of indefinite articles / number of articles	3
Proper names per noun [a]	Number of proper names / number of nouns	3

Table 13. *(continued)*

Feature name	Calculation	Ref.
Connectives of type X per sentence (Breindl) *a for X = additive*	Number of connectives of type X defined by Breindl / number of sentences; options for types of connectives for X: *causal, additive, adversative, temporal, concessive, adversative or concessive, other* Ex.: causal connectives per sentence	3
Connectives of type X per sentence (Eisenberg) *a for X = additive*	Number of connectives of type X defined by Eisenberg / number of sentences; options for types of connectives for X: see above Ex.: causal connectives per sentence	
Multi-word connectives per sentence (Breindl)	Number of multi-word connectives defined by Breindl / number of sentences	3
Multi-word connectives per sentence (Eisenberg)	Number of multi-word connectives defined by Eisenberg / number of sentences	3
Single-word connectives per sentence (Breindl) *a*	Number of single-word connectives defined by Breindl / number of sentences	3
Single-word connectives per sentence (Eisenberg) *a*	Number of single-word connectives defined by Eisenberg / number of sentences	3
Connectives per sentence (Breindl) *a*	Number of connectives defined by Breindl / number of sentences	3
Connectives per sentence (Eisenberg)	Number of connectives defined by Eisenberg / number of sentences	3
wenn-V1 conditionals per sentence	Number of conditional clause constructions using wenn-V1 / number of sentences	2
V1-dann conditionals per sentence	Number of conditional clause constructions using V1-dann / number of sentences	2
V1-V1 conditionals per sentence	Number of conditional clause constructions using V1-V1 / number of sentences	2
Coverage of conditional types	Number of conditional clause types observed at leats once / number of conditional clause types distinguished in the code (= 3; *wenn-V1, V1-dann, V1-V1*)	2
PID *a*	see Brown, Snodgrass, Kemper, Herman, and Covington (2008)	ibid.
Language use (58)		
The individual frequency data-bases used for the following features are: dlexDB (Heister et al., 2011), SUBTLEX-DE and Google Books 2000 (Brysbaert et al., 2011), and frequencies extracted from the KCT corpus (Lavalley, Berkling, & Stüker, 2015).		
Type frequency per type found in frequency data-base X *a for all X*	Sum of the frequencies of all word types in the text that were found in frequency data-base X / number of word types in the text found in frequency data-base X; frequency data-base options for X: *dlexDB, Google Books 2000,*	1, 4

Table 13. *(continued)*

Feature name	Calculation	Ref.
	SUBTLEX-DE, KCT corpus Ex.: Type frequency per type found in dlexDB	
Log type frequency per type found in frequency data-base X [a for all X]	Sum of the log frequencies of all word types in the text that were found in frequency data-base X / number of word types in the text found in frequency data-base X; frequency data-base options for X: see above Ex.: Log type frequency per type found in dlexDB	1, 4
Annotated type frequency per type found in dlexDB [a]	Sum of the frequencies of all word types in the text with their PoS annotation that were found in dlexDB / number of word types in the text that were found in dlexDB	1
Log annotated type frequency per type found in dlexDB [a]	Sum of the log frequencies of all word types in the text with their PoS annotation that were found in dlexDB / number of word types in the text that were found in dlexDB	1
Lemma frequency per type found in frequency data-base X [a, b for all X]	Sum of the frequencies of all lemmas in the text that were found in frequency data-base X / number of word types in the text that were found in frequency data-base X; frequency data-base options for X: *dlexDB, KCT corpus* Ex.: Lemma frequency per type found in dlexDB	1, 4
Log lemma frequency per type found in frequency data-base X [a, b for all X]	Sum of the log frequencies of all lemmas in the text that were found in frequency data-base X / number of word types in the text that were found in frequency data-base X; frequency data-base options for X: see above Ex.: Log lemma frequency per type found in dlexDB	1, 4
Percentage of types found in frequency data-base X [a for X ∈ {*dlexDB, KCT, subtlex*}]	Number of word types in the text that were found in frequency data-base X / number of word types in the text; frequency data-base options for X: *dlexDB, SUBTLEX-DE, KCT corpus* Ex.: Percentage of types found in dlexDB	1, 4
Percentage of types not found in frequency data-base X [a for X ∈ {*dlexDB, KCT, SUBT LEX*}]	Number of word types in the text that were not found in frequency data-base X / number of word types in the text; frequency data-base options for X: see above Ex.: Percentage of types not found in dlexDB	1, 4
Percentage of lemmas found in frequency data-base X [a, b]	Number of lemmas in the text that were found in frequency data-base X / number of word types in the text; frequency data-base options for X: *dlexDB, KCT corpus* Ex.: Lemma frequency per type found in dlexDB	4
Familiarity score per million per type in frequency data-base X [a for all X]	Cumulative frequency per million for all word types starting with the same three characters in data-base X / number of word types in the text found in data-base X; frequency data-base options for X: Google *Books 2000, SUBTLEX-DE* Ex.: Familiarity score per million per type in SUBTLEX-DE	4

Table 13. *(continued)*

Feature name	Calculation	Ref.
Log familiarity score per million per type in frequency data-base X *a for all X*	Log of cumulative frequency per million for all word types starting with the same three characters in data-base X / number of word types in the text found in data-base X; frequency data-base options for X: see above Ex.: Log familiarity score per million per type in SUBTLEX-DE	4
Log annotated type frequency band X per type found in dlexDB *a for X*∈ {3, 4, 5}	Number of annotated types in log frequency band X / number of word types in the text found in dlexDB; *frequency band options for X: integers ranging from 1 to 6* Ex.: Log annotated type frequency band 3 per type found in dlexDB	1
Log type frequency band X per type found in KCT corpus *a for X*∈ {1, 5}	Number of annotated types in log frequency band X / number of word types in the text found in KCT; frequency band options for X: integers ranging from *1* to *5* Ex.: Log annotated type frequency band 3 per type found in KCT	4
Log type frequency band X per type found in SUBTLEX-DE *a for X*∈ {2, 3, 4, 5}	Number of annotated types in log frequency band X / number of word types in the text found in SUBTLEX-DE; frequency band options for X: integers ranging from *1* to *6* Ex.: Log annotated type frequency band 3 per type found in SUBTLEX-DE	4
Log type frequency band X per type found in Google Books 2000 *a for X*∈ {4, 5, 6, 7}	Number of annotated types in log frequency band X / number of word types in the text found in Google Books 2000; frequency band options for X: integers ranging from *1* to *9* Ex.: Log annotated type frequency band 3 per type found in Google Books 2000	4
Average age of active use for word types *a*	Sum of ages of children contributing writings to the KCT corpus in which word types in text occur / number of word types in the text that were found in the KCT corpus	4
Minimal age of active use for word types *a*	Sum of ages of youngest child contributing writings to the KCT corpus in which word types in text occur / number of word types in the text that were found in the KCT corpus	4
Maximal age of active use for word types	Sum of ages of oldest child contributing writings to the KCT corpus in which word types in text occur / number of word types in the text that were found in the KCT corpus	4
Average age of active use for lemma types *a, b*	Sum of ages of children contributing writings to the KCT corpus in which lemma types in text occur / number of lemma types in the text that were found in the KCT corpus	4
Minimal age of active use for lemma types *a, b*	Sum of ages of youngest child contributing writings to the KCT corpus in which lemma types in text occur / number of lemma types in the text that were found in the KCT corpus	3

Table 13. *(continued)*

Feature name	Calculation	Ref.
Maximal age of active use for lemma types [b]	Sum of ages of oldest child contributing writings to the KCT corpus in which lemma types in text occur / number of lemma types in the text that were found in the KCT corpus	4
Human processing (25) *(Gibson, 2000; Shain et al., 2016)*		
Sum longest dependency per sentence [a, b]	Sum of number of words in the longest dependency in each sentence / number of sentences	1
Longest dependency [a, b]	Maximal number of words in a dependency in the text	1
Maximal total integration cost per finite verb using configuration X [a, b for all X]	Sum of maximal total integration costs at he finite verb calculated using the configuration X / number of finite verbs; configuration options for X: *original weights (O), increased verb weight (V), decreased coordination weight (C), decreased modifier weight (M) and weight adjustment combinations: CV, CM, VM, CMV* Ex.: Maximal total integration cost per finite verb using CV weights (= decreased coordination weights and increased verb weights)	4
Total integration cost per finite verb using configuration X [a, b for all X]	Sum of total integration costs at the finite verb calculated using configuration X / number of finite verbs; configuration options for X: see above Ex.: Total integration cost per finite verb using CV weights (= decreased coordination weights and increased verb weights)	4
Adjacent high integration costs per finite verb using configuration X [b for all X]	Sum of adjacent integration costs > 2 after a finite verb calculated using configuration X / number of finite verbs; configuration options for X: see above Ex.: Adjacent high integration cost per finite verb using CV weights (= decreased coordination weights and increased verb weights)	4
Surface text measures (3)		
Number of sentences	Total number of sentences	1
Number of paragraphs	Total number of paragraphs	1
Number of words [a]	Total number of words	1

Address for correspondence

Zarah Weiss
University of Tübingen
Department of Linguistics
Wilhelmstraße 19
72074 Tübingen
Germany
zweiss@sfs.uni-tuebingen.de

Co-author information

Detmar Meurers
University of Tübingen
Department of Linguistics
dm@sfs.uni-tuebingen.de

Modeling the Readability of German Targeting Adults and Children: An empirically broad analysis and its cross-corpus validation

Zarah Weiß **Detmar Meurers**
ICALL-Research.com
Department of Linguistics
University of Tübingen
{zweiss,dm}@sfs.uni-tuebingen.de

Abstract

We analyze two novel data sets of German educational media texts targeting adults and children. The analysis is based on 400 automatically extracted measures of linguistic complexity from a wide range of linguistic domains. We show that both data sets exhibit broad linguistic adaptation to the target audience, which generalizes across both data sets. Our most successful binary classification model for German readability robustly shows high accuracy between 89.4%–98.9% for both data sets. To our knowledge, this comprehensive German readability model is the first for which robust cross-corpus performance has been shown. The research also contributes resources for German readability assessment that are externally validated as successful for different target audiences: we compiled a new corpus of German news broadcast subtitles, the *Tagesschau/Logo* corpus, and crawled a *GEO/GEOlino* corpus substantially enlarging the data compiled by Hancke et al. (2012).

Zusammenfassung

Wir untersuchen zwei neue Datensätze deutscher Bildungs- und Mediensprache für Kinder und Erwachsene. Die Analyse basiert auf 400 automatisch extrahierten Maßen sprachlicher Komplexität, die verschiedene linguistische Domänen abdecken. Unsere Ergebnisse zeigen, dass in beiden Datensätzen die sprachliche Gestaltung der Texte in ähnlicher Weise breitflächig an ihr jeweiliges Zielpublikum angepasst wird. Unser erfolgreichstes binäres Klassifikationsmodell erzielt Genauigkeitswerte von 89,4% und 98,9% über beide Datensätze hinweg. Unseres Wissens handelt es sich bei diesem umfassend durch verschiedene linguistische Bereiche informtiertem Modell deutscher Text-Lesbarkeit um das erste, für das robuste Ergebnisse in einer korpusübergreifenden Evaluation dokumentiert sind. Darüber hinaus tragen wir mit unserer Arbeit zwei neue Datensätze zur Erforschung deutscher Text-Lesbarkeit bei, die auf Texten basieren, deren Eignung für ihre respektiven Zielgruppen extern durch wiederholte Rezeption validiert wurde: Wir haben aus Untertiteln deutscher Nachrichtenbeiträge das *Tagesschau/Logo* Korpus erstellt. Weiterhin haben wir das *GEO/GEOlino* Korpus beträchtlich erweitert, das ursprünglich von Hancke et al. (2012) erstellt wurde.

1 Introduction

Readability assessment refers to the task of (automatically) linking a text to the appropriate target audience based on its complexity. A diverse spectrum of potential application domains has been identified for this task in the literature, ranging from the design and evaluation of education materials, to information retrieval, and text simplification. Given the increasing need for learning material adapted to different audiences and the barrier-free access to information required for political and social participation, automatic readability assessment is of immediate social relevance. Accordingly, it has attracted considerable

Proceedings of the 27th International Conference on Computational Linguistics, pages 303–317
Santa Fe, New Mexico, USA, August 20-26, 2018.

research interest over the last decades, particularly for the assessment of English (Crossley et al., 2011; Chen and Meurers, 2017; Feng et al., 2010).

For German readability assessment, however, little progress has been made in recent years, despite a series of promising results published around the turn of the decade (Vor der Brück et al., 2008; Hancke et al., 2012). In particular, German readability research has suffered from the lack of a shared reference corpus and sufficiently comparable corpora for cross-corpus testing of readability models: While for English research, the *Common Core* corpus consisting of examples from the English Language Arts Standards of the Common Core State Standards, and the *WeeklyReader* corpus of online news articles have been widely used in studies on English readability and text simplification (Vajjala and Meurers, 2014; Petersen and Ostendorf, 2009; Feng et al., 2010), there are no comparable resources for German. This is particularly problematic, as over-fitting is a potential issue for classification algorithms, especially given the limited size of the typical data sets.

To address these issues, we first present two new data sets for German readability assessment in Section 3: a set of German news broadcast subtitles based on the primary German TV news outlet *Tagesschau* and the children's counterpart *Logo!*, and a GEO/GEOlino corpus crawled from the educational GEO magazine's web site, a source first identified by Hancke et al. (2012), but double in size.[1] The longstanding success of these outlets with their target audiences provides some external validity to the nature of the implicit linguistic adaptation of the language used. As Bryant et al. (2017) showed for German secondary school textbooks, this is not necessarily the case across all linguistic dimensions and adjustments may even be limited to only the surface level of text, sentence, and word length. We conducted a series of analyses on these two data sets to accomplish the following objectives:

1. Investigate how instances of German educational news language differ in terms of language complexity across adult and child target audiences.

2. Build a binary readability model for German educational language targeting adults and children that shows high, robust classification performance across corpora.

For the purposes of our studies, we operationalize child target audience of German educational news language as children aged between 8 and 14. This is the typical audience age range of the child-targeting news media we analyzed.[2] Adult target audience then is defined as over 14 years of age.

To address our first research question, after introducing a broad set of complexity measures in Section 4, we compare their informativeness for distinguishing adult and child level in the two data sets in Section 5. In Section 6, we define a series of readability models for German, including one showing high classification accuracy between 89.4% and 98.9% on both data sets. The paper closes with a discussion of the implications of our results for the current research discussion and an outlook on future work.

2 Related Work

For over a century, text readability has been assessed using surface measure-based readability formula such as the Flesch-Kincaid formula (Kincaid et al., 1975) or the Dale-Chall readability formula (Chall and Dale, 1995), see for an overview DuBay (2004). While these formula are still used in some non-linguistic studies (Woodmansey, 2010; Grootens-Wiegers et al., 2015; Esfahani et al., 2016), a decade ago research shifted towards using more elaborate statistical modeling approaches based on larger sets of linguistically more informed features. Automatic readability assessment has benefited from the use of Natural Language Processing tools for the assessment of syntactic, lexical, and discourse measures and from adapting complexity measures employed in Second Language Acquisition research (Vajjala and Meurers, 2012; Feng et al., 2010). There has also been extensive research on the relevance of cohesion

[1] We are currently negotiating with the broadcasters of *Tagesschau* and *Logo!* and the publishers of *GEO*/*GEOlino* to make the data freely available to other researchers and will make it available from `http://www.icall-research.de` in that case.

[2] The *GEOlino* magazine is advertised as targeting children between 8 and 14 years (cf. `http://www.geo.de/magazine/geolino-magazin`, accessed 11.06.18, 15:49). *Logo!* does not specify the age of its target audience, but has been reported to be particularly popular with children from age 8 to 12 (vom Orde, 2015).

and discourse measures for readability assessment that have successfully been employed for proficiency assessment in the *CohMetrix* project (Crossley et al., 2008; Crossley et al., 2011). Another example is the work by Feng et al. (2010), who evaluate which of the typically proposed measures of text readability are most promising by studying their relevance on primary school students reading material. They find language model features and cohesion in terms of entity density to be particularly useful, as well as measures of nouns. Interestingly, they also observe overall sentence length to be more informative than more elaborate syntactic features. While Feng et al. (2010) do not elaborate further on other lexical measures than POS features, Chen and Meurers (2017) conduct an elaborate cross-corpus study on the use of word frequency features for readability assessment. They show, that the typical aggregation of word frequencies across documents are less informative than richer representations including frequency standard deviations.

In contrast to English, research on readability assessment for other languages, such as German, is more limited. There was a series of articles on this issue from the late 2000s to the early 2010s that demonstrated the benefits of broad linguistic modeling, in particular the use of morphological complexity measures for languages with rich morphological systems like German (Vor der Brück et al., 2008; Hancke et al., 2012), but also Russian (Reynolds, 2016) or French (François and Fairon, 2012). The readability checker *DeLite* of Vor der Brück et al. (2008) is one of the first more sophisticated approaches that went beyond using simple readability formulas for German. The tool employs morphological, lexical, syntactical, semantic, and discourse measures, which they trained on municipal administration texts rated for their readability by humans in an online readability study involving 500 texts and 300 participant, resulting in overall 3,000 ratings. However, due to the specific nature of the data, the robustness of the approach across genres is unclear. Municipal administration language is so particular that results are unlikely to generalize to educational or literary materials, which are more attractive in first and second language acquisition contexts.

Later work by Hancke et al. (2012) also combines traditional readability formula measures, such as text or word length, with more sophisticated lexical, syntactic, and language model, and morphological features to assess German readability, but they employ an overall broader and more diverse feature set than *DeLite*. They investigate readability of educational magazines on the GEO/GEOlino data set, which they compiled from online articles freely available at the GEO magazine's web page. Their work illustrates the relevance of rich linguistic modeling for readability assessment and in particular the value of morphological complexity features for German.

The latest large scale research endeavor for the assessment of German text readability has focused more on identifying linguistic differences between texts targeting different audiences than on building readability models: In the *Reading Demands* project, complexity differences in German secondary school book texts across grade levels and school types were investigated. Berendes et al. (2017) and Bryant et al. (2017) analyze to which extent publishers successfully adapt their reading material to their target audiences. They find a lack of consistent adaptation for passive constructions, concessive and adversative connectives, and relative clauses, and only some limited adaptation in terms of lexical variation, noun complexity, and dependency length measures.

3 Data Sets

3.1 GEO/GEOlino

The GEO/GEOlino data set consists of online articles from one of the leading German monthly educational magazines, *GEO*, and the counterpart for children, *GEOlino*.[3] They are comparable to the *National Geographic* magazine and cover a variety of topics ranging from culture and history to technology and nature. Hancke et al. (2012) first compiled and analyzed a data set from this web resource. We followed their lead and crawled 8,263 articles from the GEO/GEOlino online archive, almost doubling the size of the original corpus. We removed all material flagged as non-article contents by GEO as well as all articles that contained less than 15 words. We further cleaned our data from crawling artifacts and performed near-duplicate detection with the Simhash algorithm. We then grouped all texts into topic categories

[3] `http://www.geo.de` and `http://www.geo.de/geolino`

based on the subdomains they were published under, following the web page topic structure.[4] Table 1 shows the composition of the corpus in terms of the topic groups. Since the number of documents in the different topic groups differ between GEO and the smaller GEOlino set, we created a more balanced subset (GEO/GEOlino$_S$). For this, we included only topic categories existing in both GEO and GEOlino, included all GEOlino texts in those categories and sampled from the GEO texts in those categories until we reached the same overall size of 2480 texts each.

Topic	GEO	GEOlino	$\sum$	GEO$_S$	GEOlino$_S$	$\sum_S$
Do It Yourself	0	663	663	0	0	0
Humanity	1,476	1,168	2,644	1,047	1,168	2,215
Nature	1,704	576	2,280	1,218	576	1,794
Reviews	300	736	1,036	215	736	951
Technology	0	121	121	0	0	0
Travel	1,519	0	1,519	0	0	0
$\sum$	4,999	3,264	8,263	2,480	2,480	4,960

Table 1: Distribution of topics in the full and sampled GEO/GEOlino data set.

3.2 Tagesschau/Logo

The Tagesschau/Logo data set is compiled from subtitles of German daily news broadcasts of *Tagesschau* and its children's counterpart *Logo!*. *Tagesschau* is the dominant national television news service of Germany, produced by the German public-service television network ARD. It broadcasts multiple updated editions of daily news throughout the day. *Logo!* is a television news service for children produced by the German public-service television broadcaster ZDF airing once a day. The data set consists of subtitles for all editions of both news outlets that have been broadcasted from December 2015 to January 2017. For this paper, we limited the *Tagesschau* data to the main edition broadcasted at 8pm. This amounts to overall 421 editions for *Tagesschau* and 415 editions for *Logo!*, with the small difference arising from a lack of Logo! broadcasts on some public holidays or due to special broadcasts. We cleaned the subtitles by removing non-spoken comments (e.g., * *music playing* * or * *cheering* *).

3.3 Characteristics of the two data sets

Table 2 compares the profiles of the GEO/GEOlino$_S$ and the Tagesschau/Logo data sets that we used.

	GEO$_S$	GEOlino$_S$	Tagesschau	Logo
Num. Documents (total)	2,480	2,480	421	415
Num. Words (median)	383	350	1631	1322
Num. Sentences (median)	23	25	167	125

Table 2: Corpus profile for sampled GEO/GEOlino data set and the Tagesschau/Logo data set.

While GEO/GEOlino contains more documents than Tagesschau/Logo, they are considerably shorter in terms of the number of words and sentences they contain. Another difference arises in terms of the medium: GEO/GEOlino articles are self-contained reading material and Tagesschau/Logo subtitles

[4] Subdomains were mapped to topic groups in the following way based on the URL components following `http://www.geo.de` and `http://www.geo.de/geolino`: building ("basteln"), learning ("lernen"), children's recipes ("kinderrezepte"), and competitions ("wettbewerbe") were categorized as DO IT YOURSELF. Jobs ("berufe"), extras ("extras"), photography ("fotografie"), creativity ("kreativ"), info ("info"), love ("liebe"), magazines ("magazine"), human ("mensch"), idioms ("redewendungen"), and knowledge ("wissen") were categorized as HUMANITY. Nature ("natur"), nature and environment ("natur-und-umwelt"), and animal encyclopedia ("tierlexikon") were labeled as NATURE. Book reviews ("buechertipps"), movie reviews ("filmtipps"), game reviews ("spieletest"), and GEO television ("geo-tv") were labeled as REVIEWS. Research and technology ("forschung-und-technik") was labeled as TECHNOLOGY. Travel ("reisen") was labeled as TRAVEL.

complement video material. At the same time, they consist of German educational media language and share the functional goal of conveying information to the reader, so that we consider them to be sufficiently similar to support a cross-corpus analysis.

4 Complexity Analysis

For the assessment of German language complexity, we extract 400 complexity measures using state of the art NLP techniques. All features are theoretically grounded in the contemporary research in linguistic subdisciplines, in particular Second Language Acquisition research, where Complexity is one of three dimensions of language proficiency, together with Accuracy and Fluency (Housen et al., 2012). SLA research has a rich tradition of analyzing the complexity development of learner language, see Lu (2010; 2012) for an overview. Vajjala and Meurers (2012) show that these measures can be successfully applied to readability research. Building on these findings, we follow the SLA definition of complexity as the elaborateness and variability of language (Ellis and Barkhuizen, 2005). Our measures can be grouped into seven categories: i) lexical complexity, ii) clausal complexity, iii) phrasal complexity, iv) morphological complexity, v) discourse complexity, vi) cognitive complexity, and vii) language use. While the former five groups are rooted in the linguistic system, the latter two categories were derived from psycholinguistic research. The resulting complexity assessment covers a broad variety of measures. To the best of our knowledge, this is currently the most extensive feature collection for German complexity assessment.[5] Table 3 gives an overview of the feature categories and how much they contribute to our assessment.[6]

Category	#	Description
Descriptive	2	Total number of sentences and words.
Lexical	73	Lexical diversity measures such as general and POS-specific type-token ratios as well as semantic relatedness measures.
Sentential	119	Ratios measuring sentential elaboration and variation, such as clauses per sentence.
Phrasal	41	Ratios measuring phrasal elaboration and variation, such as modifiers per noun phrase.
Morphological	39	Ratios of inflection, derivation, and composition measures.
Cohesion	48	Subsequent (local) or across text (global) use of implicit or explicit cohesion markers such as connectives, pronouns, or grammatical transitions.
Cognitive	23	Dependency lengths, verb-argument distances, and ratios of cognitive integration costs assessing cognitive processing load based on Gibson's (2000) Dependency Locality Theory.
Language Use	54	Word frequency ratios based on Subtlex-DE (Brysbaert et al., 2011), dlexDB (Heister et al., 2011), Karlsruhe Children's Texts (Lavalley et al., 2015) Approximation of age of active use based on Karlsruhe Children's Texts.

Table 3: Overview over complexity measures grouped by feature categories.

In order to extract these measures, we employ an elaborate analysis pipeline which relies on a number of NLP tools and external linguistic resources. We use OpenNLP 1.6.0 for tokenization and sentence segmentation.[7] This serves as input for the Mate tools 3.6.0 (Bohnet and Nivre, 2012), which perform a morphological analysis, lemmatization, POS tagging, and dependency parsing. We then use the JWord-

[5] Our feature collection draws from varying perspectives on language complexity including SLA and human language processing research. While the confirmation or refutation of specific theories underlying these measures is an interesting research endeavor, our empirical questions focus on which of these features support the distinction of texts targeting different audiences.

[6] We are working on integrating our German complexity analysis pipeline into CTAP (Chen and Meurers, 2016) to make it generally available and will include an online documentation for each feature.

[7] `http://opennlp.apache.org`

Splitter 3.4.0 for compound analysis.[8] The Mate POS tags are further used to inform the Stanford PCFG parser 3.6.0 (Rafferty and Manning, 2008) and the Berkeley parser 1.7.0 (Petrov and Klein, 2007), which we use for constituency and topological field parsing. For all tools, we use the German default models that were provided with them, except for the Berkeley parser, for which we use the topological field model by Ziai and Meurers (2018). With these annotations, we extract all instances of the linguistic constructs that we need to calculate the final 400 complexity ratios.[9]

5 Study 1: Which complexity measures are informative?

5.1 Set-Up

We first want to determine the informativeness of each measure for distinguishing between adult and child target audience. For this, we calculate the information gain of each measure on both data sets using 10-folds cross-validation for training and testing. We then compare across both data sets i) the number of features that are informative, and ii) the 20 most informative measures that show a Pearson correlation smaller than ± 0.8 with each other.[10] This allows us to gain insights into the range of linguistic properties of the documents targeting adults and children. We used WEKA (Hall et al., 2009) to calculate information gain and R for the correlation analysis.

5.2 Results and Discussion

Table 4 shows the percentage of measures that exhibited an average information gain above zero.

Data Set	Percentage	Informative to Total
GEO/GEOlino	79.00%	316/400
Tagesschau/Logo	88.25%	353/400

Table 4: Percentage of informative measures based on 10-folds cross-validated information gain.

Overall, 79.00% of the measures are informative for the GEO/GEOlino data and 88.25% for the Tagesschau/Logo data. This shows, that the documents are adjusted to their different target audiences in terms of a broad range of dimensions of linguistic complexity.

Table 5 provides a deeper look into the linguistic design of the documents by showing the 20 most informative measures distinguishing adult from child targeted documents, including only measures with a correlation less than ± 0.8. The table shows the original rank of each measure before removal of correlated measures, the average merit of each measure for the distinction of the target audience, the type of complexity measures it belongs to, and the feature name.

The results for both data sets show a diverse collection of features, some of which are similar for both data sets, but also some interesting differences. In total the measures seem to be more informative for Tagesschau/Logo, as indicated by the higher average merit, and more correlated, as can be seen from the wider range of original ranks. Language use as captured by frequency measures is particularly relevant for both data sets. The table includes seven measures of word frequency for GEO/GEOlino and five for Tagesschau/Logo. For both data sets, the most informative measure is one of language use: For GEO/GEOlino it is the average minimal age of active use of lexical types found in the Karlsruhe Children's Corpus (KCT) of Lavalley et al. (2015). For Tagesschau/Logo it is the average log lexical type frequency based on Google Books 2000. The other language-use measures are very similar across data sets: Lexical types unknown to the Subtlex-DE data base (Brysbaert et al., 2011), for example, rank 4th and 2nd on both data sets and while on Tagesschau/Logo the lemma frequency per lexical type found in KCT is the 12th most informative measure, its log counterpart ranks 8th on GEO/GEOlino.

[8] `http://www.danielnaber.de/jwordsplitter`

[9] To support transparent comparison with other complexity studies, we include a description of the operationalization of all linguistic units that allow for varying definitions in Appendix A, as has been suggested by Bulté and Housen (2014).

[10] We set the Pearson correlation threshold relatively high since we primarily are interested in qualitatively inspecting the types of measures that are informative, not in removing all correlations.

GEO/GEOlino				Tagesschau/Logo			
Rank	Average Merit	Group	Feature	Rank	Average Merit	Group	Feature
1	0.332 (±0.004)	USE	sumTypesMinAoAPerTypeInKCT	1	0.978 (±0.004)	USE	logTypeFreqsPerTypeInGoogle00
4	0.327 (±0.005)	LEX	syllablesPerToken	31	0.899 (±0.006)	USE	typesNotInSubtlexPerLexicalType
11	0.288 (±0.004)	USE	logTypeFreqsPerTypeInSubtlex	50	0.825 (±0.009)	COH	2PPersPronounsPerNoun
13	0.231 (±0.003)	USE	typesNotInSubtlexPerLexicalType	71	0.754 (±0.012)	COH	probNotSubsPerTransition
15	0.205 (±0.003)	COH	2PPersAndPossPronounsPerToken	82	0.716 (±0.010)	COH	causalConnectivePerSentence
21	0.164 (±0.004)	USE	typesNotInDlexPerLexicalType	85	0.689 (±0.009)	COH	localArgOverlapsPerSentence
23	0.147 (±0.004)	PHR	complexNominalsPerTUnit	88	0.667 (±0.008)	SEN	sumParseTreeHeightsPerFiniteClause
24	0.143 (±0.002)	USE	logLemmaFreqsPerTypeInKCT	89	0.662 (±0.008)	SEN	NPsPerTUnit
25	0.143 (±0.003)	SEN	syllablesInMiddleFieldPerMiddleField	90	0.656 (±0.007)	COH	1PPersPronounsPerToken
28	0.133 (±0.003)	COH	persPronounsPerToken	91	0.657 (±0.010)	MOR	genitivesPerNoun
31	0.133 (±0.003)	SEN	PPsPerTUnit	95	0.633 (±0.011)	PHR	determinersPerNP
33	0.132 (±0.003)	MOR	secondPersonMarkingsPerFiniteVerb	100	0.622 (±0.011)	USE	lemmaFreqsPerTypeInKCT
35	0.123 (±0.004)	MOR	ionTPerToken	101	0.620 (±0.014)	COG	sumLongestDependenciesPerClause
36	0.122 (±0.003)	LEX	synsetPerTypeInGnet	102	0.617 (±0.008)	MOR	compundNounsPerNP
37	0.121 (±0.004)	PHR	complexNominalsPerFiniteClause	103	0.609 (±0.012)	USE	typeFreqsPerTypeInDlex
38	0.120 (±0.004)	SEN	sumNonTerminalNodesPerTUnit	109	0.568 (±0.008)	LEX	MTLD
40	0.118 (±0.004)	USE	typeFreqsPerTypeInSubtlex	110	0.560 (±0.010)	LEX	nonAuxVerbTypesPerNonAuxVerbToken
43	0.114 (±0.002)	COH	pronounsPerNoun	111	0.550 (±0.013)	COH	globalStemOverlapsPerSentence
44	0.113 (±0.002)	USE	logAnnoTypeFreqBand5PerTypeInKCT	117	0.505 (±0.011)	SEN	conjunctionalClausesPerSentence
49	0.111 (±0.002)	COH	3PPersAndPossPronounsPerNoun	119	0.500 (±0.014)	USE	logAnnoTypeFreqBd4PerTypeInDlex

Table 5: Top 20 most informative measures on balanced GEO/GEOlino and Tagesschau/Logo data based on information gain with $r \leq 0.8$.

Cohesion measures are highly informative, too, although more so for Tagesschau/Logo. In particular the use of certain personal or possessive pronouns is highly informative for GEO/GEOlino. The use of second person pronouns ranks highly for both data sets, which may easily be explained by it being used for the informal German address appropriate when speaking to children. This is further corroborated by the ratio of second person verb inflections being ranked as the 13th most important measure. For Tagesschau/Logo, other implicit measures of textual cohesion based on content overlap are also informative as well as the use of causal connectives. Overall 55% of the most informative 20 measures for both data set are captured by these two categories.

The other feature groups are less frequently represented, but provide some interpretable insights into the data. First, both data sets show indications of differences in the degree of nominalization used in language targeting adults and children: For GEO/GEOlino, complex noun phrases per t-unit and finite clause are highly informative as well as the use of the nominalization suffix *-ion*. On Tagesschau/Logo, genitive case, determiners per noun phrase, and the percentage of compound nouns indicate a similar relevance of differences regarding the organisation of the nominal domain. Lexical and sentential complexity seems to be less homogeneous for the distinction of adult and child targeted language across data sets: There are two measures of lexical complexity assessing word length in syllables and the semantic inter-relatedness of words ranked high for GEO/GEOlino, while on Tagesschau/Logo, lexical diversity and verb variation are particularly informative. For sentential complexity, constituency tree complexity, the average length of the middle field, and the use of prepositional phrases per t-unit are particularly informative on GEO/GEOlino. On Tagesschau/Logo, parse tree height and the use of conjunctional clauses are relevant. Cognitive measures do not seem to play an important role on either data set, except for the sum of longest dependencies per clause on Tagesschau/Logo.

Overall, these results clearly show that for both data sets the distinction between target audiences is not just made based on surface modifications such as sentence or word length. In fact, these measures do not occur among the most informative measures at all. Rather, measures of language use and cohesion are predominantly informative for the distinction of adult and child targeting texts, but also measures of phrasal, sentential, lexical, and morphological complexity. The adjustment of the data to their audience observed here thus seems to be more linguistically refined than that found in the *ReadingDemands* textbook data, where Berendes et al. (2017) found only few adjustment across dimensions.

6 Study 2: Can we successfully model readability for German, also across data sets?

6.1 Set-Up

Our second objective is the design of a robust model of educational media language that distinguishes robustly between language targeting adults and children across corpora and genres. For this, we train two binary Sequential Minimal Optimization (SMO) support vector classifier (Platt, 1998) with linear kernels using the WEKA machine learning toolkit (Hall et al., 2009). Each model is tested i) on the same corpus it is trained on, using 10-folds cross-validation, and afterwards ii) on the other data set for cross-corpus testing after training on the full data set. For model performance evaluation, we report classification accuracy and the classification confusion matrices, and random baselines as reference point.

6.2 Results and Discussion

Table 6 shows the accuracy of our SMO models on both data sets and compares them with a random baseline. Both models clearly outperform the baseline of 50.0%. On GEO/GEOlino$_S$, the performance is comparable to the performance observed by Hancke et al. (2012) on the original GEO/GEOlino data.[11]

As Table 7a shows, erroneous classifications are roughly balanced across both classes, showing that the model does not prefer one class over the other. When training a model using only the 20 most informative measures identified in Study 1, we reach an accuracy of 85.1%, i.e., the additional measures only account only for 3.3%.[12] When testing the models on the Tagesschau/Logo corpus, accuracy increases to 98.8% for both models. The confusion matrix for the model using 400 measures in Table 7b seems to indicate

[11] After observing these results, we obtained the original GEO/GEOlino data set from Hancke et al. (2012) and trained and tested a model with 10-folds cross-validation on it. When using the same data, our model outperforms their best performing

Model	Training	Testing	Features	Accuracy	SD
Baseline		GEO/GEOlino$_S$		50.0	
		Tagesschau/Logo		50.0	
10-folds CV	GEO/GEOlino$_S$	GEO/GEOlino$_S$	400	89.4	$\pm$0.09
			20	85.1	$\pm$0.09
	Tagesschau/Logo	Tagesschau/Logo	400	99.9	$\pm$0.04
			20	99.8	$\pm$0.03
Cross-Corpus	GEO/GEOlino$_S$	Tagesschau/Logo	400	98.9	
			20	98.8	
	Tagesschau/Logo	GEO/GEOlino$_S$	400	52.2	
			20	56.7	

Table 6: Classification performance of model on GEO/GEOlino$_S$ and Tagesschau/Logo data

$\downarrow$Obs./Prd.$\rightarrow$	Child$_{GEOlino}$	Adult$_{GEO}$
Child$_{GEOlino}$	2,222	258
Adult$_{GEO}$	267	2,213

(a) 10-folds CV on GEO/GEOlino$_S$

$\downarrow$Obs./Prd.$\rightarrow$	Child$_{GEOlino}$	Adult$_{GEO}$
Child$_{Logo!}$	408	7
Adult$_{TS}$	2	419

(b) Cross-corpus testing on Tagesschau/Logo

Table 7: Confusion matrices for testing models with 400 features trained on GEO/GEOlino$_S$.

a minor tendency towards classifying *Logo!* texts as Tagesschau texts, but due to the low number of incorrect classifications this is not conclusive.

Overall, performance of both models trained on GEO/GEOlino$_S$ on the Tagesschau/Logo data is comparable to the performance of both models trained and tested on Tagesschau/Logo with 10-folds cross-validation, although the confusion matrix for the cross-validated Tagesschau/Logo model using 400 measures does not exhibit any tendency towards predicting one class preferred over the other, as may be seen in Table 8a.

$\downarrow$Obs./Prd.$\rightarrow$	Child$_{Logo!}$	Adult$_{TS}$
Child$_{Logo!}$	415	0
Adult$_{TS}$	1	420

(a) 10-folds CV on Tagesschau/Logo

$\downarrow$Obs./Prd.$\rightarrow$	Child$_{Logo!}$	Adult$_{TS}$
Child$_{GEOlino}$	2,472	8
Adult$_{GEO}$	2,362	118

(b) Cross-corpus testing on GEO/GEOlino$_S$

Table 8: Confusion matrices for testing models with 400 features trained on Tagesschau/Logo

The model trained and tested on Tagesschau/Logo reaches an unexpectedly high accuracy of 99.9% for using 400 measures and 99.8% when using only the 20 most informative measures reported in Study 1. Since the performance remains high when using only 20 measures and the standard deviation across folds is very low, the results seem not to be due to over-fitting. The model learns linguistic properties of the data set that generalize across. It is important to stress here than none of our measures include n-gram language models or any other lexical content features but only complexity measures aggregated over each document.[13]

model with 91.1%, confirming that our approach is in fact competitive with the state of the art.

[12]We do not show the confusion matrices for the models with 20 features, because they are equivalent to the matrices in Table 7. The same holds for the models tested on Tagesschau/Logo and their matrices in Table 8.

[13]Content features are problematic since they can pick up recurring phrases that are characteristic of particular media outlets rather than generalizable linguistic complexity characteristics. E.g., the *Tagesschau* always starts with the greeting "Hier ist das Erste Deutsche Fernsehen mit der Tagesschau." (*Here is the first public German TV channel with the daily news.*).

When testing the models trained on the Tagesschau/Logo data set on the GEO/GEOlino$_S$ data, it becomes apparent that the characteristics learned from the Tagesschau/Logo data set do not generalize, with the model based on 400 measures performing only marginally above chance, and the model using the 20 measures performing slightly better with 56.2%. When considering the confusion matrix for this model in Table 8b, we see that most texts are classified as GEOlino texts, irrespective of whether they belong to GEO or GEOlino. The Tagesschau/Logo trained models do not generalize well to the other adult/child corpus. Since the model trained on GEO/GEOlino$_S$ is highly successful when tested on Tagesschau/Logo, this cannot be due to an actual lack of generalizable differences in the linguistic characteristics of the adult and child targeting texts contained in both data sets. One possible reason for these results may be that, as Study 1 showed, the measures are considerably more informative on Tagesschau/Logo than on GEO/GEOlino$_S$. It could be, that the differences between the news subtitles designed for different target audiences are more extreme than those observable for the GEO magazines. This would explain the surprisingly good performance of the GEO/GEOlino$_S$ model on the Tagesschau/Logo data, which would then be easier to distinguish, while also accounting for the poor performance in the opposite case.

7 Summary and Outlook

We presented a study of the difference between German targeting adults and children, as far as we know the most broadly based linguistic complexity analysis to date. We created and analyzed a novel data set compiled from German news subtitles that consists of news broadcasts for adults and children from the same days, ensuring a relatively parallel selection of topics. We compared this with a newly compiled GEO/GEOlino corpus consisting of online articles of two magazines for adults and children by the same publisher discussing the similar topics. Based on these two data sets, we presented within-corpus (10-fold CV) and cross-corpus experiments and built binary classification models of German educational media text readability that perform with very high accuracy across both data sets. The model is based on a broad range of features that are highly informative for both data sets. This model is a valuable contribution since i) it is based on a considerably broader data basis than previous approaches to German readability, and ii) it successfully generalizes across the data sets, illustrating surprising robustness across rather different text types. The approach presented thus extends the state-of-the-art in Hancke et al. (2012) in terms of the breadth of features integrated and the accuracy and generalizability of the model – and provides two new data sources for this line of research.

The paper also contributes some new insights into the linguistic characteristics of German media language targeting adults and children. Since all the language is produced by adults, it is not necessarily clear how well it is in fact adjusted to the target audience. As demonstrated by Berendes et al. (2017), German textbook publishers indeed do not seem to be adjusting the complexity of the language used according to school type and grade level in any systematic way. Our results for educational media language indicate, that i) both data sets are successfully and broadly adapted towards their target audiences; and ii) that they form two distinct, cross-corpus generalizable constructs of German educational media language for children and adults. In a next step, we plan to test to which extent this linguistically diverse and generalized construct matches the language competence of the intended children target group by comparing it with the Karlsruhe Children's Text corpus (Lavalley et al., 2015). We also plan to further investigate the linguistic properties of our two data sets. In particular, the Tagesschau/Logo data set requires further statistical and qualitative analyses to investigate why its linguistic characteristics generalize well across all folds of the data set itself but not across GEO/GEOlino. We also plan to conduct more analyses of the informativeness of the different complexity feature groups for the target audience distinction.

Acknowledgements

We would like to thank Peter Lindner of the NDR and the editorial office of *ARD Aktuell* for providing us with the news subtitles of the *Tagesschau* broadcasts and Christiane Müller of the ZDF for giving us access to the subtitles from *Logo!*.

References

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pußel, Marco Rower, Bettina Schrader, Anne Schwartz, George Smith, and Hans Uszkoreit, 2003. *TIGER Annotationsschema*. Universität des Saarlandes and Universität Stuttgart and Universität Potsdam.

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2017. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics, Association for Computational Linguistics.

Doreen Bryant, Karin Berendes, Detmar Meurers, and Zarah Weiß. 2017. Schulbuchtexte der Sekundarstufe auf dem linguistischen Prüfstand. Analyse der bildungs- und fachsprachlichen Komplexität in Abhängigkeit von Schultyp und Jahrgangsstufe. In Mathilde Hennig, editor, *Linguistische Komplexität – ein Phantom?* Stauffenburg Verlag, Tübingen.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in german. *Experimental Psychology*, 58:412–424.

Bram Bulté and Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26(0):42 – 65. Comparing perspectives on L2 writing: Multiple analyses of a common corpus.

Jeanne S. Chall and Edgar Dale. 1995. *Readability revisited: the new Dale-Chall Readability Formula*. Brookline Books.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, Osaka, Japan, December. The International Committee on Computational Linguistics.

Xiaobin Chen and Detmar Meurers. 2017. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*. JRIR-2017-01-0006.R1.

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara, 2008. *Assessing text readability using cognitively based indices*, pages 475–493. Teachers of English to Speakers of Other Languages, Inc. 700 South Washington Street Suite 200, Alexandria, VA 22314.

Scott A. Crossley, David B. Allen, and Danielle McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1):84–101, April.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information, Costa Mesa, California.

Duden. 2009. *Deutsche Grammatik*, volume 4. Dudenverlag, 4 edition.

Rod Ellis and Gary Barkhuizen. 2005. *Analysing learner language*. Oxford University Press.

B. Janghorban Esfahani, A. Faron, K. S. Roth, P. P. Grimminger, and J. C. Luers. 2016. Systematic readability analysis of medical texts on websites of german university clinics for general and abdominal surgery. *Zentralblatt fur Chirurgie*, 141(6):639–644.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.*

Thomas François and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.

Petronella Grootens-Wiegers, Martine C. De Vries, Tessa E. Vossen, and Jos M. Van den Broek. 2015. Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, 37(1):89–117.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Julian Heister, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken, and Reinhold Kliegl. 2011. dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, 62:10–20.

Alexis Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Kellogg W. Hunt. 1970. Do sentences in the second language grow like those in the first? *TESOL Quarterly*, 4(3):195–202.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Rémi Lavalley, Kay Berkling, and Sebastian Stüker. 2015. Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Languages Journal*, pages 190–208.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, PaGe '08, pages 40–46, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marga Reis. 2001. Bilden Modalverben im Deutschen eine syntaktische Klasse? *Modalität und Modalverben im Deutschen*.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA, June. Association for Computational Linguistics.

Karsten Schmidt. 2016. Der graphematische Satz. *Zeitschrift für germanistische Linguistik*, 44(2):215–256.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Joel Tetreault, Jill Burstein, and Claudial Leacock, editors, *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Montréal, Canada, June. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297, Gothenburg, Sweden, April. ACL, Association for Computational Linguistics.

Heike vom Orde. 2015. Kindernachrichten im Fernsehen. Eine Zusammenfassung zentraler Forschungsergebnisse zum Format logo! *Televizion*, 28/2015/2:40–42.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429–435.

Karl Woodmansey. 2010. Readability of educational materials for endodontic patients. *Journal of Endodontics*, 36:1703–1706.

Ramon Ziai and Detmar Meurers. 2018. Automatic focus annotation: Bringing formal pragmatics alive in analyzing the Information Structure of authentic data. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, LA. ACL. To appear.

Appendix A. Definitions of Linguistic Units

Clauses are all maximal projections of finite verbs and elliptical constructions with sentential status (i.e. all sub-trees tagged with *S*), as well as *to* infinitives that have a sentential status (*satzwertige zu Infinitive*).

Complex t-units are t-units that include subordinate clauses.

Conjunctional clauses are all dependent clauses that are introduced by a subordinating conjunction such as *dass*, *weil*, or *wenn*.

Dependent clauses with conjunction are all conjunctional clauses, but also interrogative and relative clauses. Dependent clauses without conjunction are mostly dependent main clauses, such as *Ich weiß, es ist spät*.

(Graphematic) sentences are strings of at least one token that are ended by sentence ending punctuation marks: *!*, ., *?*. There is a broad discussion on alternative sentence definitions, see for example Schmidt (2016) for a more elaborate theoretical account. However, since sentences are identified by sentence segmentation tools, which are primarily based on punctuation, sentences are always defined as graphematic sentences.

Half modals are *haben*, *sein*, *scheinen*, *drohen*, *versprechen*, if they govern an infinitive with *zu* (Duden, 2009, §101), e.g. *ist zu machen*, *droht zu schneien*.

Lexical words are all nouns, adjectives, adverbs, foreign words, numbers, main verbs, and modal verbs. Note that there is an ongoing discussion on whether modals actually qualify as lexical words (Reis, 2001), hence there is also a subset of lexical words excluding modals employed throughout the system.

Parts-of-Speech are operationalized following the Tiger POS tags (Albert et al., 2003, 121).

Quasi passives are *bekommen*, *erhalten* or *kriegen* if they govern a past participle (Duden, 2009, §179), e.g. *bekommt gemacht*, *kriegt eröffnet*.

T-units are "one main clause plus any subordinate clause or non clausal structure that is attached to or embedded in it" (Hunt, 1970, 4).

Appendix B. Example Extracts from *Tagesschau* and *Logo!* subtitles

Report on New Years shooting in Istanbul by *Tagesschau*, extracted from the subtitles for the broadcast on 01.01.2017, 20:00.

> In der Türkei ist der Jahreswechsel von einem Anschlag in Istanbul überschattet: Mind. ein bewaffneter Angreifer drang in einen Nachtclub ein und schoss um sich. 39 Menschen wurden getötet und mehr als 60 verletzt. Unter den Todesopfern sind zahlreiche Ausländer. Ob Deutsche betroffen sind, ist unklar. Die Suche nach dem Täter dauert an, bekannt hat sich niemand. Das Attentat ereignete sich im europäischen Teil Istanbuls. Dort liegt direkt am Bosporus der Club "Reina", der bei Prominenten beliebt ist. Nur eine Stunde währte in der Türkei die Hoffnung, 2017 könnte ruhiger werden als 2016, das von Bombenanschlägen geprägt war. Doch um 1.15 Uhr Ortszeit macht im Istanbuler Nachtclub "Reina" ein Attentäter mit einem Gewehr Jagd auf Gäste einer Silvesterparty. Zuvor wurde vorm Club ein Polizist erschossen. Der Täter konnte fliehen, eine Großfahndung läuft. Bis zu 800 Personen sollen sich in der Diskothek aufgehalten haben. Gäste berichten, Panik sei ausgebrochen. Einige Besucher sollen in den Bosporus gesprungen sein. Unter den Toten und Verletzten sind Ausländer. Bekannt hat sich niemand zu der Tat. Türkische Medien vermuten den IS hinter dem Terrorakt. Die Regierung verhängte eine Nachrichtensperre. Wir lassen uns vom Terror nicht beirren. Was hier passierte, kann morgen an einem anderem Ort geschehen. Es gibt keine Garantien. Der Nachtclub "Reina" liegt am Bosporus, im Stadtteil Ortaköy. Er ist der berühmteste der Türkei, teuer und bei Touristen beliebt. Die Sicherheitsvorkehrungen waren landesweit erhöht worden. In Istanbul waren 17.000 Polizisten im Einsatz. Trotz Großaufgebot der Polizei, hochaktiver Geheimdienste, Ausnahmezustand und markiger Politikerworte: Die Sicherheitslage in der Türkei spitzt sich zu. Beängstigende Aussichten für Wirtschaft und Menschen.

Report on New Years shooting in Istanbul by *Logo!*, extracted from the subtitles for the broadcast on 02.01.2017, 19:50 (no broadcast on 01.01.2017).

> In der türkischen Großstadt Istanbul hat es an Silvester einen Anschlag in einer Disco gegeben. Ein Mann stürmte mit einem Gewehr in den Club und hat 39 Menschen getötet, darunter auch zwei Männer, die in Deutschland gelebt haben. Die türkische Polizei sucht jetzt nach dem Täter. Er ist seit dem Anschlag auf der Flucht. Auch am zweiten Tag nach dem Anschlag kamen viele Menschen an die Polizeiabsperrung, um Blumen für die Opfer niederzulegen. Der Terrorist stürmte dort in der Silvesternacht mit einem Gewehr in die Disco. Ich war völlig geschockt, konnte mich nicht bewegen. Der Täter schoss erst auf einen Polizisten und dann auf die Gäste. Wir hörten plötzlich Schüsse, da sind wir raus aus dem Ballsaal auf die Terrasse und haben uns dort versteckt. Im Internet behauptet die Terrorgruppe IS, Islamischer Staat, dass sie hinter dem Anschlag stecke. Die Kämpfer dieser Terrorgruppe wollen, dass alle Menschen nach ihren strengen religiösen Regeln leben. Wer sich nicht daran hält, wird sogar umgebracht. Besonders aktiv ist der IS in Teilen von Syrien und dem Irak. Beide Länder grenzen an die Türkei. Dort haben die Kämpfer in letzter Zeit schon öfter Anschläge verübt. In der ganzen Türkei sucht die Polizei jetzt nach dem Attentäter. Acht Verdächtige wurden schon festgenommen. Auf logo.de könnt ihr mehr zur Terrorgruppe Islamischer Staat lesen und da gibt es auch viele Infos zu unserem nächsten Thema.

Appendix C. Example Articles from GEO and GEOlino

GEO article titled "Was ist ein Planet?" (What is a Planet?).[14] It discusses criteria celestial bodies need to fulfill to be considered a planet.

> Lange bezeichneten Menschen alle Lichtpunkte, die über den Nachthimmel wanderten, als Planeten (griech. planáomai = umherirren) – gleich, ob es sich um Venus, Mars, Mond oder Asteroiden handelte. In der Neuzeit durften den Titel nur noch die großen Himmelskörper tragen, die um die Sonne kreisten, aber keine Monde waren – also nicht ihrerseits einen anderen Planeten umrundeten. Als Astronomen von 1992 an in den Randbezirken des Sonnensystems immer neue Objekte entdeckten, manche ähnlich groß wie Pluto (bis dahin der neunte Planet), sah sich die Internationale Astronomische Union genötigt, erstmals zu definieren, was ein Planet genau ist. Nach heftigen Diskussionen beschlossen die Astronomen 2006 die Resolution B5. Demnach muss ein Planet drei Kriterien erfüllen: Er muss um die Sonne kreisen. Er muss ausreichend Masse aufweisen, sodass er unter seiner eigenen Schwerkraft eine nahezu runde Form angenommen hat. Und er muss die Umgebung seiner Umlaufbahn freigeräumt haben. Objekte, die ihm auf seiner Bahn nahekommen, "schluckt" er in einer Kollision oder schleudert sie in einen anderen Orbit. Pluto, Eris und andere große Himmelskörper zählen nun zu den Zwergplaneten, da sie es nicht schaffen, ihre Bahn zu bereinigen, sondern sie sich mit anderen Objekten teilen. Damit kreisen nach derzeitigem Stand acht Planeten um die Sonne. Die Astronomen unterteilen sie in die vier terrestrischen Planeten Merkur, Venus, Erde, Mars (sie werden wegen ihrer festen Oberfläche häufig steinige Planeten genannt) und in die vier jovianischen – jupiterähnlichen – Planeten Jupiter, Saturn, Uranus, Neptun (aufgrund ihrer Zusammensetzung oft als Gasplaneten oder Gasriesen bezeichnet). Wobei Uranus und Neptun manchmal auch als "Eisriesen" beschrieben werden, da sie weniger Wasserstoff als Jupiter und Saturn enthalten, dafür mehr gefrorenes Methan, Wasser und Ammoniak.

GEOlino article titled "Sieben erdähnliche Planeten entdeckt" (Seven Earth-Like Planets Discovered).[15] It reports on the discovery of seven new planets that orbit Trappist-1.

> Dass neue Planeten entdeckt werden, ist erstmal nichts ungewöhnliches. Doch der Fund dieser sieben sogenannten Exoplaneten (Planeten wie Kepler-452b , die sich um einen Stern - außerhalb des Einflusses unserer Sonne - bewegen) ist etwas ganz Besonderes: Denn sechs der neu entdeckten Planeten liegen in einer Temperaturzone, in der Leben möglich ist. Auf den meisten Planeten ist es entweder kochend heiß oder eiskalt - schwierige Bedingungen für die Entwicklung von Leben. Die Sonne der Exoplaneten, der Zwergstern Trappist-1 , ist viel kleiner als die Sonne unseres Sonnensystems : Trappist-1 besitzt nur acht Prozent der Masse unserer Sonne und zwölf Prozent ihres Durchmessers. Auf drei der entdeckten Exoplaneten könnte sogar Wasser existieren, denn ihr Abstand zur Zwergsonne liegt in einem Temperaturbereich, in dem Wasser weder gefrieren noch verdampfen würde. Hier wäre also eine Art von Leben möglich, wie wir es auf unserer Erde kennen.
>
> Die sieben Planeten haben in etwa die Größe unserer Erde und sind wahrscheinlich Gesteinsplaneten. Sie alle umkreisen ihre Sonne, den Stern Trappist-1, der knappe 40 Lichtjahre (ein Lichtjahr ist die Strecke, die Licht in einem Jahr zurücklegt) von uns entfernt im Sternenbild Wassermann liegt. Weil die Sonne des Trappist-1-Systems so klein ist, können die Planeten diese wesentlich schneller umkreisen als wie es in unserem Sonnensystem möglich ist. Die sechs Planeten, die dem Zwergstern am nächsten sind, umrunden ihn in eineinhalb bis zwölf Tagen. Sie haben damit einen engeren Orbit als der Merkur um die Sonne .

[14] `http://www.geo.de/wissen/weltall/15396-rtkl-definitionssache-was-ist-ein-planet`, accessed 11.06.18, 16:06.

[15] `http://www.geo.de/geolino/wissen/weltraum/sieben-erdaehnliche-planeten-entdeckt`, last accessed 11.06.18, 16:06.

Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment

Zarah Weiss, Xiaobin Chen, Detmar Meurers
Department of Linguistics & LEAD Graduate School
University of Tübingen
Germany
{zweiss,xchen,dm}@sfs.uni-tuebingen.de

Abstract

We investigate the readability classification of English and German reading materials for language learners based on a broad linguistic complexity feature set supporting the parallel analysis of both German and English. After illustrating the quality of the feature set by showing that it yields state-of-the-art classification performance for the established OneStopEnglish corpus (Vajjala and Lučić, 2018), we introduce the Spotlight corpus. This new data set contains graded reading materials produced by the same publisher for English and German, which supports an analysis comparing the linguistic characteristics of texts at different reading levels across languages. As far as we are aware, this is both the first readability corpus for German L2 learners, as well as the first corpus with comparably classified reading material for learners across multiple languages.

After discussing the first results for a readability classifier for German L2 learners, we show that the linguistic complexity analyses for the cross-language experiments identify features successfully characterizing the readability of texts for language learners across languages, as well as some language-specific characteristics of different reading levels.

1 Introduction

The language input available to language learners is a driving force for Second Language Acquisition (SLA), and reading is an important source of language input. Material that is just above the level of the learner is assumed to be best for fostering learning, which depending on the SLA tradition is characterized as i+1 input of Krashen (1981), input in the Zone of Proximal Development in socio-cultural approaches (Lantolf et al., 2015), or input reflecting second language development in usage-based SLA approaches (Ellis and Collins, 2009). Note that the focus here is not just on input that is understandable and of interest to the learner but also rich in developmentally proximal language properties.

This dependency of readability on reading purpose and individual language skills makes the identification of appropriate reading materials a major challenge for educators, especially for heterogeneous learning groups. Automatic readability assessment may facilitate the retrieval of appropriate reading materials for individual language learners. It refers to the task of identifying texts that are suitable for a given group of target readers with a specific reading purpose (Collins-Thompson, 2014). Recent approaches to automatic readability assessment also investigate the use of neural networks (Martinc et al., 2019). However, the identification of linguistic characteristics that impact the readability of texts in itself can also yield valuable insights for education, because it may inform content creators of reading materials for language learning. This also is an interesting research endeavor from a linguistic perspective and speaks against solely focusing on neural approaches. Similarly, it remains to be investigated to which extent these linguistic characteristics may generalize across languages given comparable target groups and reading purposes.

While there has been a considerable amount of work on automatic readability assessment for English, there is still insufficient research on other

Zarah Weiss, Xiaobin Chen and Detmar Meurers 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings 177: 38–54.

languages. The lack of suitable training corpora for other languages remains as one major limiting factor (Collins-Thompson, 2014), despite some research efforts to facilitate unsupervised readability assessments (Benzahra and François, 2019; Martinc et al., 2019). For example, there has been some recent work on German readability classifiers for native speakers (Weiss and Meurers, 2018; Weiss et al., 2018; Dittrich et al., 2019). Yet, a lack of corpus resources has so far hindered the development of a readability classifier for German as a second or foreign language (L2) learners.

In this article, we introduce a novel cross-lingual feature collection for broad linguistic modeling of German and English complexity. Although neural classification approaches have been strongly represented in readability assessment, our literature review (see Section 2) shows that their success has been very much limited on the benchmark data we use for this study and fallen behind the feature-based readability classification approaches which are also providing deeper linguistic insights while requiring less computational power.[1] However, while broad feature collections for language-specific complexity modeling have been proposed for English (Chen and Meurers, 2019) and German (Weiss and Meurers, 2018), they are not applicable across languages. This has so far hindered the cross-lingual study of similarities between characteristics of readability. We first validate our approach by applying it to an established readability corpus for English (Vajjala and Lučić, 2018), before using it to train two readability classifiers for labeling English and German L2 reading materials resulting in the first readability classifier of this kind for German. For this, we introduce a novel data set of English and German reading materials for beginning, intermediate, and advanced learners of English and German, the Spotlight corpus. We address the following research questions:

1. Can we train a successful readability classifier for German and for English using broad complexity modeling?

2. Can these classifiers generalize beyond their training language to cross-lingual contexts?

3. Which linguistic features are relevant for the distinction of reading levels and how do they differ between English and German?

The article is structured as follows. First, we discuss related work on readability assessment of English and German (Section 2). Then, we introduce the novel Spotlight data set (Section 3.1) as well as the OneStopEnglish corpus (Section 3.2) which we use as benchmark data set. We proceed to introduce our approach to automatic complexity assessment and the feature set (Section 4) we use throughout our machine learning experiments (Sections 5 and 6). Finally, we compare the informativeness of individual complexity features on Spotlight for the discrimination of reading levels (Section 7) before we come to the conclusion (Section 8) and outlook (Section 9).

2 Related Work

Automatic readability assessment has a long history dating back to the first readability formulas developed in the early 20th century, see DuBay (2006) for an overview. Traditional readability formulas employ few surface text characteristics such as text, sentence, and word length (Flesch, 1948; Dale and Chall, 1948). They are still widely used especially in non-linguistic studies on web accessibility (Esfahani et al., 2016; Grootens-Wiegers et al., 2015), in information retrieval systems (Miltsakaki and Troutt, 2007; Chinkina et al., 2016), and for confirming the compliance of reading materials with specific accessibility guidelines (Weiss et al., 2018; Yaneva et al., 2016), such as Easy-to-Read materials.[2]

Over the last two decades, there has been a shift towards computational readability classification approaches based on machine learning techniques employing feature engineering with Natural Language Processing (NLP) methods, see Collins-Thompson (2014) and Benjamin (2012) for an overview. Among others, linguistic complexity features from SLA research (Vajjala and Meurers, 2012), word frequency measures (Chen and Meurers, 2017), and features of text cohesion (Crossley et al., 2017) from Writing Quality Assessment research (Crossley, 2020) were shown to be valuable features for readabilty assessemnt.

While most readability research focuses on English (Collins-Thompson, 2014), to a lesser degree these approaches have also been employed for other languages such as Russian (Reynolds, 2016),

[1] See Strubell et al. (2019) for a discussion of the considerable energy demands of deep learning approaches in NLP.

[2] https://www.inclusion-europe.eu/easy-to-read/

French (François and Fairon, 2012), Swedish (Pilán et al., 2015), Italian (Dell'Orletta et al., 2013), or German (Vor der Brück and Hartrumpf, 2007). For German, the most recent classification approach has been proposed by Weiss and Meurers (2018) who use broad linguistic complexity modeling of German to distinguish between German media texts targeting adults and children. However, this approach only provides a rather coarse binary distinction and identifies reading materials for information retrieval (i.e., with a focus on accessibility), rather than language learning (i.e., with a focus on challenging the reader's language competence). Given the lack of appropriate multi-level reading corpora, so far no classifiers for German L2 readers have been trained.

Recently, several neural network approaches have been proposed for readability assessment (Martinc et al., 2019; Madrazo Azpiazu and Pera, 2019). Martinc et al. (2019) investigate the performance of supervised and unsupervised neural readability classification approaches for English and Slovenian. They find that their neural approaches perform overall at the state-of-the art level of feature-based classification approaches in both languages. For the OneStopEnglish corpus, their best classifier reaches an accuracy of 78.71% which performs at the same level as the feature-based classifier reported by Vajjala and Lučić (2018) with an accuracy of 78.12%. With this, the performance of neural approaches on OneStopEnglish does not exceed the original benchmark and lies substantially below the current state-of-the art on this data set, which is held by a feature-based classifier with an accuracy of 90.09% (Bengoetxea et al., 2020). In other words, while neural classification approaches have been very successful in several NLP tasks, they are currently not competitive with the breadth and depth of analyses supported by feature-based approaches to readability classification.

Only little research has been conducted on multilingual readability classification. While there are some neural classification approaches that are developed to be applicable across languages (Martinc et al., 2019; Madrazo Azpiazu and Pera, 2019), feature-based approaches are usually language-specific. An exception is the study by De Clercq and Hoste (2016), who compare the informativeness of lexical, semantic and syntactic features for English and Dutch readability classification. The cross-lingual applicability of multilingual models has so far not been investigated, except for a series of studies by Madrazo Azpiazu and Pera on the VikiWiki corpus, which distinguishes simplified Vikidia.org texts for 8 to 13 year old children from regular Wikipedia.org texts for Basque, Catalan, Dutch, English, French, Italian, and Spanish.[3] On this data, Madrazo Azpiazu and Pera (2020a) investigate the transferablility of the neural readability classification approach by Madrazo Azpiazu and Pera (2019). They demonstrate that training on multilingual data sets may improve readability classification results for low-resource languages in the binary classification task. Madrazo Azpiazu and Pera (2020b) follow a similar approach using a feature-based readability classification approach based on shallow features, morphological features, syntactic features, and semantic features. They report similar results as Madrazo Azpiazu and Pera (2020a). While these studies make an important first contribution to the assessment of cross-lingual readability assessment, they are clearly limited by the binary distinction of simplified texts for children and regular Wikipedia texts. The success of transfer learning for more fine-grained and practically relevant readability level distinctions remains to be empirically determined.

3 Data

3.1 Spotlight corpus

The Spotlight corpus consists of articles from the two monthly language learning magazines *Spotlight*[4] for adult German learners of English and *Deutsch perfekt*[5] for adult language learners of German. Both magazines are published by *Spotlight Verlag*, a leading European publisher for foreign language learning materials.[6] The magazines contain reading materials for beginning, intermediate, and advanced language learners which the publisher equates with the Common European Framework of Reference (CEFR) levels A2 (level: easy), B1/B2 (level: medium) and C1 (level: advanced).

We extracted all articles from the PDF versions of the respective issues provided to us for research purposes by the publisher.The type setting of the magazines made it impossible to di-

[3] https://github.com/ionmadrazo/VikiWiki
[4] https://www.spotlight-online.de
[5] https://www.deutsch-perfekt.com
[6] https://www.spotlight-verlag.de

rectly extract the individual articles with a PDF converter without loosing the information of their reading level. Instead, we manually identified and extracted each article using screenshots which we then converted to plain text using Google's optical character recognition (OCR) API.[7] This way, we extracted the English subset (henceforth Spotlight-EN) from the 110 issues of the *Spotlight* magazine that were published from January 2012 to December 2019 and the German subset (henceforth Spotlight-DE) from the 45 issues of the *Deutsch perfekt* magazine published from January 2018 to December 2019 (see corpus profiles in Table 1). The imbalance of readability levels in both data

Level	N. docs	N. sents	N. words
Spotlight-EN			
Easy	1.030	13.921	212.267
Medium	1.528	60.232	898.695
Advanced	1.030	24.288	440.793
$\sum$	3.285	98.441	1.551.755
Spotlight-DE			
Easy	763	16.135	180.178
Medium	509	27.107	338.553
Advanced	174	11.713	155.160
$\sum$	1.446	54.955	673.891

Table 1: Corpus profiles for Spotlight data

sets is due to the imbalanced distribution of reading levels in both magazines.

It is noteworthy that in both magazines, articles may vary considerably in length irrespective of their reading level. This is shown in Table 2. The table showcases that number of words – which has been and continues to be a popular surface feature for readability classification – is not sufficient to distinguish reading levels in this data set.

3.2 OneStopEnglish corpus

The OneStopEnglish (OSE) corpus by Vajjala and Lučić (2018) consists of overall 567 Guardian news paper articles that were rewritten for adult English as a Second Language learners by MacMillan Education.[8] Each Guardian article is available in an elementary (ele), intermediate (int), and advanced (adv) version resulting in a perfectly

[7]https://cloud.google.com/vision

[8]https://www.onestopenglish.com

	$\mu \pm SD$	M	Min	Max
Spotlight-EN				
Easy	206±166	137	53	877
Medium	588±555	493	23	4.497
Advanced	606±509	489	26	2.940
Spotlight-DE				
Easy	236±235	137	60	1.469
Medium	665±769	448	72	5.605
Advanced	892±537	524	91	4.161

Table 2: Article length in words in Spotlight data ($\mu \pm SD$= mean ± standard deviation; M = median; Min = minimal; Max = maximal)

balanced corpus.[9] The OSE corpus is a by now established reference data set for studies related to readability assessment and text simplification (Bengoetxea et al., 2020; Benzahra and François, 2019). Currently, the best results reported for OSE achieve an accuracy of 90.09% in a feature-based machine learning approach by Bengoetxea et al. (2020). Table 3 shows the corpus profile of the OSE data set. Table 4 displays the differences of article length across reading levels in OSE.[10]

Level	N. docs	N. sents	N. words
Ele.	189	6.033	105.169
Int.	189	6.634	128.335
Adv.	189	7.221	162.449
$\sum$	567	19.888	395.953

Table 3: Corpus profile for OSE

Level	$\mu(\pm SD)$	M	Min	Max
Ele.	556(±109)	561	267	948
Int.	679(±117)	691	315	1.083
Adv.	860(±171)	857	357	1.465

Table 4: Article length in words in OSE ($\mu \pm SD$= mean ± standard deviation; M = median; Min = minimal; Max = maximal)

[9]Since the three OneStopEnglish levels (elementary, intermediate, advanced) are not explicitly aligned with the CEFR levels, used to characterize the Spotlight levels (easy=A2, medium=B, advanced=C1), we keep the labels separate throughout the article.

[10]The numbers reported here slightly deviate from those reported by Vajjala and Lučić (2018), due to minor differences in the automatic tokenization.

As also noted by Vajjala and Lučić (2018, p. 299), there is a general tendency of articles becoming longer with increasing reading level. However, note the standard deviation of the article length within reading levels, which is considerable despite being much lower than the variability displayed in the Spotlight data.

4 Automatic Complexity Analysis

4.1 Complexity Features

We calculate 312 features of linguistic complexity merging the feature collections proposed by us in our previous work on German (Weiss and Meurers, 2018) and English (Chen, 2018). These have been successfully used for the tasks of readability assessment (Chen and Meurers, 2018; Weiss and Meurers, 2018; Kühberger et al., 2019), second language proficiency assessment (Weiss and Meurers, 2019b, 2021), academic language proficiency (Weiss and Meurers, 2019a), and teachers' grading objectivity (Weiss et al., 2019). While each of the feature collections contains more language-specific features than the joined feature collection proposed in this work, this is as far as we are aware the broadest collection of complexity features applicable to both, English and German, thus facilitating cross-lingual comparisons of complexity.

Our broad set of cross-lingual complexity features covers the theoretical linguistic domains of syntax, lexicon, and morphology, as well as features of discourse cohesion and psycho-linguistic features of human language use and human language processing. It also includes some surface measures from or inspired by classic readability formulas.

4.1.1 Surface Length (LEN)

We measure 21 surface text length features inspired by traditional readability formulas. They measure the raw number of sentences, syllables, letters, (unique) words including and excluding punctuation marks and numbers, and (unique) tokens. It also includes mean and standard deviations of sentence length and word length measured in letters, syllables, and words as well as the mean and standard deviation of words with more than two syllables. These categories can be applied without language-specific adjustments, except for the identification of syllables which are based on language-specific regular expressions.

4.1.2 Syntactic Complexity (SYN)

We assess several features of clausal and phrasal complexity that have been proposed in the SLA complexity literature (Wolfe-Quintero et al., 1998; Kyle, 2016) inspired by the implementations by Chen (2018) and Weiss and Meurers (2021). We measure 20 features of clausal elaborateness. This includes features measuring the length of clauses and (complex) t-units in various units (such as words, syllables, letters), as well as features of clausal coordination and subordination, such as the number of relative or dependent clauses per clause.

Furthermore, we measure 28 features of phrasal elaborateness. This includes several features focusing on the complexity of noun phrases (NPs) including the number of pre- and postnominal modifiers per complex NP, the number of (complex) NPs per clause, t-unit and sentence, and the length of NPs in words. It also entails features measuring the complexity of verb phrases (VPs) including the number of verb clusters and VPs per clause, t-unit and sentence and the length of verb clusters in words. We also measure the complexity of prepositional phrases (PPs) such as the number of (complex) PPs per clause, t-unit and sentence or the length of PPs in words. Finally, this includes measures of coordinate phrases per clause, t-unit and sentence.

While these syntactic features are identified based on language-specific TregEx (Levy and Andrew, 2006) patterns for constituency trees, we carefully designed all extraction rules to yield equivalent results across languages.

We also measure syntactic variation based on 12 measures of parse tree edit distances following Chen (2018).

4.1.3 Lexical Complexity (LEX)

We measure several complexity features assessing lexical richness, variation, and density that have been proposed in the SLA complexity literature (Wolfe-Quintero et al., 1998) inspired by the implementations by Chen (2018) and Weiss and Meurers (2021). These can be applied straight forward across languages as long as similar word categories (such as adjectives, nouns, verbs, etc.) can be identified.

This feature set includes 27 features of lexical density including POS-based lexical density features as well as 9 features of lexical diversity including lexical word, verb, noun, adjective, and

adverb variation. Finally, we assess 53 features of lexical richness including several mathematical transformations of type token ratios (TTR), parts-of-speech specific TTRs, the Uber index and HD-D (McCarthy and Jarvis, 2007).

4.1.4 Morphological Complexity (MOR)

Morphological complexity has been argued to be an important feature for readability assessment of morphological richer languages than English, such as German (Hancke et al., 2012; Weiss and Meurers, 2018) or Basque (Gonzalez-Dios et al., 2014). However, few measures have been used in readability assessment that are applicable across languages with different morphological systems. We use the Morphological Complexity Index (MCI) proposed by Brezina and Pallotti (2019) to assess morphological complexity independent of language by measuring the variability of morphological exponents of specific parts-of-speech within a text. These morphological exponents can be identified by contrasting word forms with their stems which makes the features applicable across languages. We assess overall 40 MCI features for verbs, nouns, and adjectives based on different number of samples and sampling sizes with and without repetition.

4.1.5 Discourse Cohesion (DIS)

We assess 26 features measuring the mean overlap of word forms and lemmas of lexical words, nouns, and grammatical arguments between sentences as well as their standard deviation. Each feature is calculated locally (between neighboring sentences) and globally (across all sentences in the text). These implicit cohesion features were originally proposed in CohMetrix (McNamara et al., 2014). Unlike explicit cohesion measures, such as the number of particular connectives, they are directly applicable across languages.

4.1.6 Language Use (USE)

Word frequency features have a long tradition in both, readability and complexity research. Yet, word frequencies obtained from different frequency data bases are not necessarily comparable. We address this issue by using the SUBTLEX-US (Brysbaert et al., 2011b) and SUBTLEX-DE (Brysbaert et al., 2011a) frequency data bases. We consider both SUBTLEX frequency data bases equivalent for the purposes of our complexity analysis because they represent word frequencies from the same register and were created to be maximally comparable. To mitigate effects due to the different sizes of the underlying corpora, we only use word frequencies per million words.

Based on this, we calculate 56 word frequency features including the mean (log) frequency of all words, lexical words, and function words and their standard deviations as well as frequencies for verbs, nouns, adjectives, and adverbs.

4.1.7 Human Language Processing (HLP)

Weiss and Meurers (2018) have proposed to use features based on theories explaining human sentence processing difficulties for readability assessment. They propose features based on the Dependency Locality Theory (Gibson, 2000) using the different integration cost weight configurations proposed in Shain et al. (2016). While the psycholinguistic theories have been formulated for English, the complexity features by Weiss and Meurers (2018) have so far not been applied for complexity modeling beyond German.

We implemented 21 features for both, English and German, based on universal dependencies to make them applicable across languages. These features calculate the average, maximal and highest adjacent discourse integration costs per finite verb across different weight configurations.

4.2 NLP Pipeline

We calculate our complexity features following a three-step procedure. First, we run a pipeline of Natural Language Processing (NLP) tools to provide linguistic annotations for the data. The annotation pipeline primarily relies on Stanford CoreNLP (Manning et al., 2014) which we use for sentence segmentation, tokenization, parts-of-speech (POS) tagging, constituency parsing, and dependency parsing for English and German. We additionally employ the Mate tools (Bohnet and Nivre, 2012) for lemmatization, because CoreNLP only provides a lemmatizer for English but not for German. We also use the OpenNLP Snowball stemmer to extract stems for English and German. For all annotations, we use the respective default models provided with the NLP tools.

Second, we count linguistic constructs using a set of extraction rules as well as word frequencies. This procedure is fully identical across languages except for syllable counts, POS-based counts, and syntactic complexity counts which we designed to be comparable across languages as described in

the previous section. For all other features we use identical extraction rules.

Third, we calculate a variety of complexity feature ratios based on these counts. This step is fully language independent.

4.3 Feature Extraction and Selection

We extracted all 312 features on OSE, Spotlight-EN and Spotlight-DE as described in the previous subsection. We then identified all features that were not variable on any of the three data sets. This way, we could exclude features that are irrelevant for the data sets while keeping the feature collections comparable across data sets. For this, we removed all features for which the most common feature value across all three data sets occurred in 95% of the data or more.

The feature removal reduced the entire feature collection to 301 features. Only human language processing features were removed through this step, including all features measuring high adjacent integration costs.

5 Establishing our Approach on OSE

5.1 Set-up

To validate the performance of our feature-based readability classification approach against an established benchmark data set, we first trained a classifier to predict reading levels on the OSE data. For this, we used the 301 complexity features from Section 4.3. All feature values were z-transformed and centered around zero. We trained a random forest (RF), an ordinal RF, a Support Vector Machine (SVM) with a radial kernel, and a SVM with a polynomial kernel in R (R Core Team, 2015) using the `caret` package (Kuhn, 2020).[11] In the following, we only report the results for the SVM using a polynomial kernel, which outperformed the other algorithms.[12]

To not reduce the relatively small data set further, we train and test using 10-folds cross-validation. We compare the performance of the classifier on OSE with a) the random accuracy baseline of 33.3% and b) the state-of-the art performance on this data set by Bengoetxea et al. (2020), reaching 90.09%. We also report the individual precision, recall and F1 scores for each reading level.

5.2 Results

The OSE classifier reaches an accuracy of 92.06% with a 95% confidence interval (CI) $= [89.52\%, 94.15\%]$ in 10-folds cross-validation. This significantly outperforms the random baseline of 33.33% (p-Value $< 2 \cdot 10^{-16}$).[13] It also exceeds the results of Bengoetxea et al. (2020).

Table 5 displays the confusion matrix for the classification summed across all 10-folds.

Pred\Obs.	Ele.	Int.	Adv.
Ele.	179	9	4
Int.	9	173	15
Adv.	1	7	170

Table 5: Confusion matrix: OSE 10-CV

It shows that misclassifications occur predominantly at adjacent reading levels and that there does not seem to be any systematic bias. Table 6 reports precision, recall, and F1 score per level. The performance across reading levels is relatively

	Ele.	Int.	Adv.
Precision	93.2	87.8	95.5
Recall	94.7	91.5	90.0
F1	94.0	89.6	92.6

Table 6: Performance for OSE 10-CV

balanced. Elementary texts have a slightly higher recall, while advanced texts have a higher precision. As expected when comparing an ordinal classification level with two adjacent levels with levels with only one adjacent level, intermediate texts receive the lowest scores for precision and recall.

6 Classifying Readability on Spotlight

6.1 Set-up

After establishing the performance of our approach against the OSE benchmark data set, we turn to our main research question, which compares feature-based readability classification across languages on Spotlight-EN for English and Spotlight-DE for German. Our classification is

[11] All R scripts, data tables, and trained models that are being reported in this and the following sections are publicly available on OSF at https://osf.io/5hbcs/

[12] SVM parameters: degree = 3, scale = 0.001, and C = 1.

[13] Here and throughout the article we report p-values obtained with one-sided t-tests with $H_1 = Acc. > Baseline$.

again based on the 301 complexity features we extracted and identified following the procedure described in Section 4.3. All feature values were z-transformed and centered around zero separately for Spotlight-EN and Spotlight-DE. This way, the classifiers are learning based on the standard deviations from the data sets' mean values rather than the raw feature values. This was supposed to mitigate language-specific differences, for example, regarding the average sentence length in German and English.

The set-up of the classification experiment is identical to the one described in Section 5.1. In the following, we only report the results for the ordinal RF which outperformed the other algorithms on both Spotlight data sets.[14] Since this is a novel data set, we use the majority baseline as sole reference to evaluate the classifier performance in the within language condition (Section 6.2.1).

For our cross-language classification experiment (Section 6.2.2), we apply the previously trained classifiers to the respective other subset of the Spotlight data, i.e., testing on Spotlight-DE for the classifier trained on Spotlight-EN and vice versa. Unlike previous cross-linguistic readability classification approaches that used cross-lingual data to augment limited training resources, this set-up tests the generalization of our classifiers in a form of zero-shot learning. We again compare the performance of each classifier across-languages against the majority baseline on the respective testing data and the within-language classification performance.

We also report the individual precision, recall and F1 scores for each reading level throughout all classification experiments.

6.2 Results

6.2.1 Within-language Performance

Table 7 displays the results of all four classification experiments on the Spotlight data. The Spotlight-EN classifier reaches an accuracy of 74.5% in 10-folds cross-validation. This significantly outperforms the majority baseline of 46.5% (p-Value $< 2.2 \cdot 10^{-16}$).

Looking at the confusion matrix in Table 8, we see that the classification is relatively balanced, even though in proportion to their total count advanced texts are classified incorrectly more often than the other reading levels. This can also be seen in the relatively low F1 score for advanced texts displayed in the first three rows of Table 10.

The Spotlight-DE classifier reaches an accuracy of 88.0% in 10-folds cross-validation. This significantly outperforms the majority baseline of 52.8% (p-Value $< 2.2 \cdot 10^{-16}$). Table 9 shows the confusion matrix for the classification, which shows good classification results throughout all reading levels. This is mirrored in the high precision and recall scores displayed in rows four to six in Table 10.

6.2.2 Cross-language Performance

For the classification across languages, the Spotlight-EN classifier reaches an accuracy of 55.5% on Spotlight-DE. Although this performance is considerably worse than for the within-language classification, this significantly outperforms the majority baseline of 52.8% (p-Value = 0.02118) showing that the classifier somewhat generalizes beyond English even if the performance drops considerably. Looking at the confusion matrix in Table 11, one of the most common misclassifications is the labeling of easy texts as medium. The classifier overestimates the reading difficulty of many easy and medium texts. This results in a high precision but low recall for easy texts, as shown in rows seven to nine in Table 10.

The Spotlight-DE classifier reaches an accuracy of 53.4% on Spotlight-EN. Again, this is much worse than the results for the within-language classification, but significantly outperforms the majority baseline of 46.51% (p-Value $= 1.284 \cdot 10^{-15}$). This shows again that the classifier generalizes to some degree in the zero-shot learning scenario. Looking at the confusion matrix in Table 12, it can be seen that the classifier tends to underestimate the reading difficulty of advanced texts (classifying them as medium or even easy) and of medium texts (classifying them as easy). This results in a relatively high recall for easy texts and very low recall for advanced texts, as shown in the final three rows in Table 10.

6.3 Discussion

The two readability classifiers trained on Spotlight-EN and Spotlight-DE are highly successful when applied within their training language and exceed the majority baseline con-

[14] Parameters for the English model: number of sets = 50, number of trees per div. = 150, number of final trees = 600; parameters for the German model: number of sets = 150, number of trees per div. = 150, number of final trees = 200.

Train	Test	Acc.	95% CI	Maj.	Acc. $<$ Maj.
Spotlight-EN	10-folds CV	74.5	$[73.0, 76.0]$	46.5	$< 2.2 \cdot 10^{-16}$
Spotlight-DE	10-folds CV	88.0	$[86.1, 89.6]$	52.8	$< 2.2 \cdot 10^{-16}$
Spotlight-EN	Spotlight-DE	55.5	$[52.9, 58.1]$	52.8	.02118
Spotlight-DE	Spotlight-EN	53.4	$[51.7, 55.1]$	46.5	$1.284 \cdot 10^{-15}$

Table 7: Overall classifier accuracy (Acc.) on Spotlight data compared against majority baseline (Maj.)

Pred\Obs.	Easy	Medium	Advanced
Easy	816	171	37
Medium	208	1,210	268
Advanced	6	147	422

Table 8: Confusion matrix Spotlight-EN 10-CV

Pred\Obs.	Easy	Medium	Advanced
Easy	727	83	1
Medium	34	399	27
Advanced	2	27	146

Table 9: Confusion matrix Spotlight-DE 10-CV

	Easy	Medium	Advanced
Spotlight-EN 10 CV			
Precision	79.7	71.8	73.4
Recall	79.2	79.2	58.1
F1.	79.5	75.3	65.0
Spotlight-DE 10 CV			
Precision	89.6	86.7	83.4
Recall	95.3	78.4	83.9
F1.	92.4	82.4	83.7
Spotlight-EN on Spotlight-DE			
Precision	82.3	42.5	52.4
Recall	44.6	67.4	67.8
F1.	57.8	52.1	59.2
Spotlight-DE on Spotlight-EN			
Precision	49.3	59.0	53.4
Recall	80.3	47.9	27.0
F1.	61.1	52.9	35.8

Table 10: Level-wise performance on Spotlight

siderably. When comparing the performance of the Spotlight-EN classifier and the OSE classifier, the different nature of the two English corpora has to be taken into account. OSE consists of the same 189 articles simplified for three different reading levels, which is a somewhat artificial set-up for training data. The Spotlight-EN corpus, instead, consists of different texts specifically written for a given reading level which is closer to real-life texts for which language learners might require automatic readability ratings. Thus, we consider the within-language performance of the Spotlight-EN classifier satisfactory.

Pred\Obs.	Easy	Medium	Advanced
Easy	341	73	0
Medium	408	343	56
Advanced	14	93	118

Table 11: Confusion matrix Spotlight-EN on Spotlight-DE

Pred\Obs.	Easy	Medium	Advanced
Easy	827	635	216
Medium	193	732	315
Advanced	10	161	196

Table 12: Confusion matrix Spotlight-DE on Spotlight-EN

For the Spotlight-DE classifier, we observe a very high performance throughout reading levels. Spotlight-DE is the first data set for the readability assessment of texts for German L2 learners that allows a distinction for beginning, intermediate, and advanced learners of German. Thus, we cannot compare the performance to a reference corpus or cross-corpus test the Spotlight-DE classifier. Overall, the classification results are sufficient to use the Spotlight-DE classifier in real-life scenarios, even though a cross-corpus evaluation on a comparable data set would be ideal to confirm its generalizability as soon as such a data set becomes available.

Turning to our cross-language classification experiments, we find that both classifiers generalize

to some extent in the zero-shot learning scenarios, despite considerable drops in performance. This result is not to be taken for granted due to the linguistic differences between English and German. These are highly promising initial results. Further research is needed to investigate to which extent this generalization also applies across other languages.

The comparison of the confusion matrices of both cross-lingual classification experiments reveals a symmetrical regularity in the misclassifications. While the German classifier underestimates the reading levels of the English texts, the English classifier tends to overestimate the readability of the German texts. Since the classifiers are trained and tested on feature z-scores centered around the mean this behavior is not immediately expected and warrants further investigation in future research.

7 Feature Informativeness on Spotlight

7.1 Set-up

To identify which of the 301 complexity features identified in Section 4.3 are most informative for the readability classification, we identify the most informative features using the correlation-based feature subset selection for machine learning approach by Hall (1999). This method identifies the subset of features that exhibits the highest correlation with the class to be predicted (in our case reading level) while minimizing the intercorrelation of features within the subset. We use the implementation provided in the WEKA toolkit version 3.9.5 (Hall et al., 2009) for feature identification. We report the percentage of features selected across each feature group before we discuss in more detail the intersection of features in both data sets.

7.2 Results

Table 13 displayed the raw number and percentage of features selected on Spotlight-EN and Spotlight-DE across feature groups and the total number of features contained in the feature group. To make the result summary more interpretable, we split syntactic and lexical complexity features into the individual subgroups distinguished within Sections 4.1.2 and 4.1.3. A full list of all features that are informative on either data set is displayed in Appendix A. Figure 1 shows the boxplots of all features that were selected for Spotlight-EN as well as Spotlight-DE.

Group	EN (%)		DE (%)		All
LEN	7	(33.3)	5	(23.8)	21
USE	17	(30.4)	11	(19.6)	56
LEX Density	7	(15.9)	5	(18.5)	27
LEX Diversity	1	(11.1)	1	(11.1)	9
LEX Richness	4	(7.5)	5	(9.4)	53
SYN Clausal	1	(5.0)	8	(40.0)	20
SYN Phrasal	1	(3.6)	5	(17.9)	28
SYN Variation	2	(16.7)	0	(0.0)	12
MOR	7	(17.5)	3	(7.5)	40
DIS	2	(8.2)	0	(0.0)	24
HLP	0	(0.0)	0	(0.0)	11
$\sum$	49	(16.3)	43	(14.3)	301

Table 13: Informative features selected on Spotlight-EN (EN), Spotlight-DE (DE), and the total number of features in the group (All)

On Spotlight-EN and on Spotlight-DE, up to a third of all surface length features are selected, most of which are informative on both data sets. All of the shared length features increase with reading level (see Figure 1). Also language use features seem to be central for the distinction of reading levels on both data sets. 30.4% of the features were selected for Spotlight-EN and 19.6% for Spotlight-DE. Four of the language use features are relevant for both data sets: the average word frequency and its standard deviation are decreasing with increasing reading level. The same holds for the log frequency of lexical word types. The standard deviation of the verb token frequency is increasing with higher reading levels. Lexical complexity seems to play a medium role in the distinction of reading levels. 13.5% of the lexical complexity features were selected for Spotlight-EN and 12.4% for Spotlight-DE. Especially lexical density and richness play an important role on both data sets, but there is only very little overlap between the features selected for Spotlight-EN and Spotlight-DE. Only the POS density of modifiers and proper nouns as well as the squared word TTR were selected on both feature sets. For English, the proper noun density is decreasing, while the POS density for modifiers and the squared word TTR are increasing with reading levels. For German, the squared word TTR is also increasing with reading levels, but the two POS density features exhibit a u-shaped and inverse u-shaped

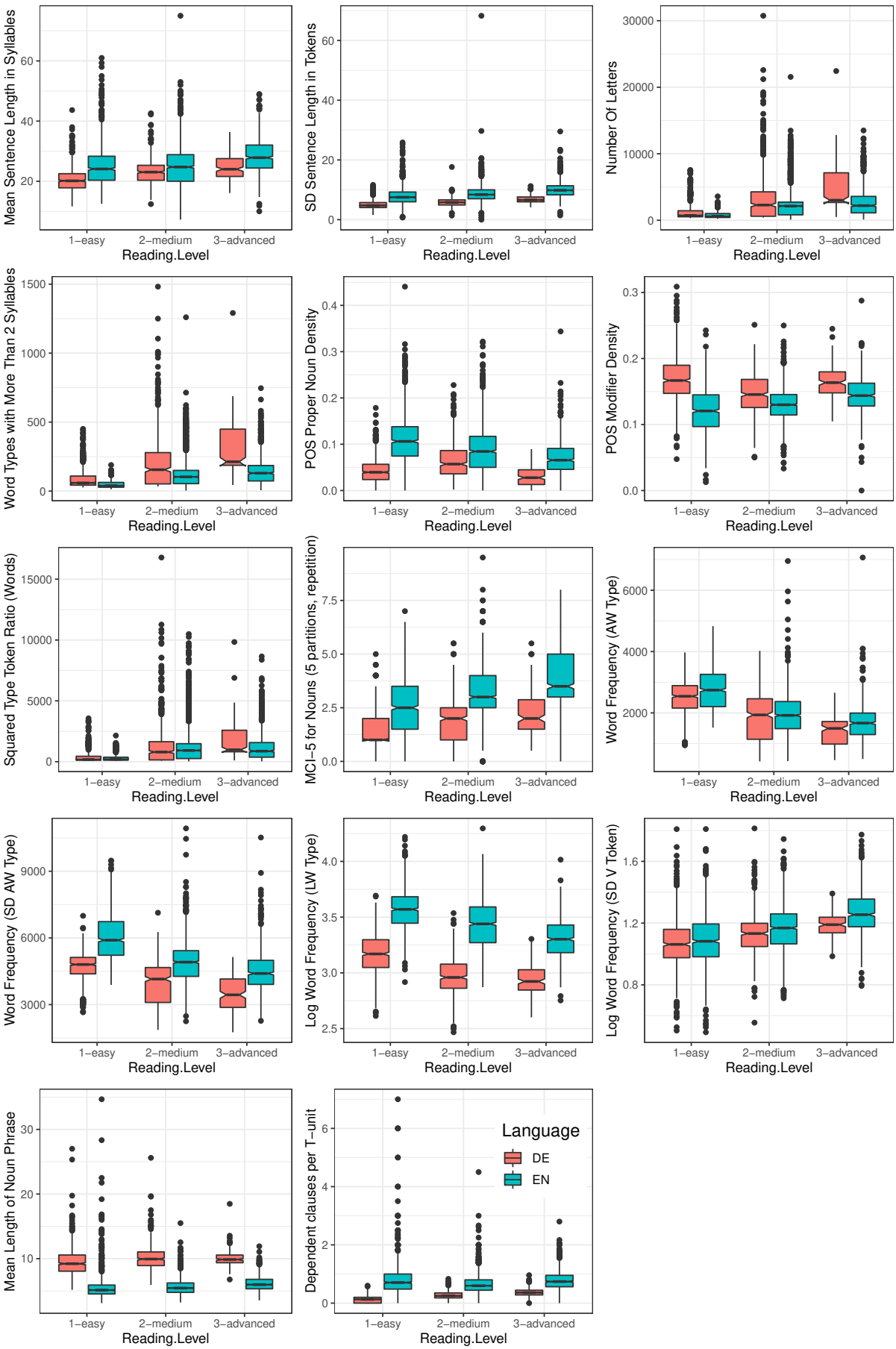


Figure 1: Boxplots of features that are informative on both, Spotlight-EN and Spotlight-DE

behavior.

The importance of syntactic and morphological complexity differs for Spotlight-EN and Spotlight-DE. Only 6.7% of the syntactic features were selected for Spotlight-EN, half of them features of syntactic variation. In contrast, 21.7% were selected on Spotlight-DE, all either features of clausal or phrasal complexity. Correspondingly, there is very little overlap in this domain between English an German. Only two syntactic features are informative for both data sets: the mean noun phrase length and the number of dependent clauses per t-unit, both of which are increasing with higher reading levels on both data sets. Morphological complexity features seem to play an important role for the distinction of reading levels on Spotlight-EN, but much less on Spotlight-DE. While 17.5% of the morphological complexity features were selected for Spotlight-EN, only 7.5% play a role on Spotight-DE. Both data sets share only one feature in this domain, namely the MCI for adjectives (measured with repetition with 5 partitions of size 5), which increases with higher reading levels, though the effect is more pronounced for English.

Neither implicit discourse cohesion features nor human language processing features seem to be important features on Spotlight-DE and also on Spotlight-EN, only 8.2% of the cohesion features were identified as informative.

7.3 Discussion

The correlation-based feature subset selection shows that features from most feature groups contribute meaningful information for the distinction of reading levels on both data sets. Especially features of surface length, language use, and lexical complexity help to characterize reading level differences on both data sets. Morphological and syntactic complexity features seem to capture more language-specific differences. There is also a considerable overlap of features selected for both data sets. Overall 28% of the features selected for Spotlight-EN and 32% of features selected for Spotlight-DE are shared between both data sets.

Judging from the features that are shared between the feature selections for English and German, higher reading levels are characterized by the use of less frequent vocabulary, longer words, sentences, and texts, and shifts in lexical density and richness. Also the features that were selected from the domains of morphological, phrasal and syntactic complexity increase with higher reading levels. This is in line with previous findings by Weiss and Meurers (2018) regarding the readability of German media texts targeting German-native speaking adults and children. However, our results indicate that these domains play a much less pronounced role for the distinction of reading levels. Interestingly, morphological elaboration seems to be more important for English than for German.

Human language processing measures do not seem to play an important role for the distinction of reading levels in either data sets, even though these measures are motivated by psycho-linguistic studies on human sentence processing. This is again in line with previous findings reported by Weiss and Meurers (2018).

Overall, these findings explain the albeit limited cross-language generalization of both readability classifiers in the zero-shot learning experiments. While there are differences in the types of features that are informative for the identification of reading levels across languages, there is nevertheless a substantial overlap and the shared features predominantly exhibit an increase in complexity with higher reading levels. This confirms that the publisher successfully instituted a policy facilitating the creation of stratified reading materials for different levels in a way that is comparable across the different languages that we analyzed.

8 Conclusion

We have investigated the use of language-independent broad linguistic complexity modeling for the multi-level readability classification of English and German reading materials for language learners. Our first study designed to benchmark the performance of our methods on the established OneStopEnglish yielded new state-of-the art results, clearly showcasing the value of broad linguistic modeling for readability assessment. Our study also shows that for certain tasks, broadly linguistically informed feature-based approaches are in fact not only competitive with neural approaches but exceeding their performance.

We then introduced a novel multi-level reading corpus for English and German on which we trained two readability classifiers that yield are highly successful within their respective training language. With this, we present the first multi-level readability classifier for German. This is highly relevant, because the much more com-

monly proposed binary classification approaches distinguishing simple and regular language are too limited to be of practical relevance for the retrieval of reading materials that are appropriate to foster foreign language learning.

We then demonstrated the generalizability of the German classifier for comparable English data and the English classifier for comparable German data. This is a novel contribution to cross-lingual readability research, not only because of the multi-level classification but also because of we propose a zero-shot cross-lingual readability classification approach unlike previous work focusing on augmenting low-resource training data. This is a central contribution to readability classification research, especially for languages other than English, given the lack of appropriate training materials for many languages.

In our final study, we compared the linguistic properties characterizing differences in reading levels in English and German. Our findings show that for both languages, texts systematically differ between reading levels in terms of the frequency and lexical complexity. Language-specific characteristics of reading levels can be found in the syntactic, discourse and morphological domains. The publisher thus successfully adapts the reading materials for different proficiency levels across a variety of linguistic domains in a systematic way. This is not a trivial insight, since previous work demonstrated that school book publishers do not always succeed in the linguistic adaptation of reading materials for different target groups (Berendes et al., 2018).

Our findings clearly demonstrate the value of feature-based classification approaches not only for the study of linguistic phenomena but also for readability classification. We demonstrate the feasibility of broad language-independent feature collections and their potential for zero-shot cross-lingual learning.

9 Outlook

As we saw in Table 7, cross-language zero-shot learning showed a promising result for training on Spotlight-DE and test on Spotlight-EN and the other way round. It is arguable that although different languages may complexify in different linguistic aspects, the general rule of more elaborate linguistic components and more varied expression usually resulting in higher complexity still applies. As a result, it is highly likely that zero-shot cross-language learning would also result in good performance, but detailed approaches need to be further designed and tested in future studies including more languages.

Another direction for future research is to see how the readability levels decided by the publisher match L2 learners' actual perception of the texts' difficulty. Although our models have yielded high accuracy, if the standards used to determine the levels of the texts do not actually match the learners' perceived difficulty, the predicted results are meaningless. Vajjala and Lučić (2019) offer an interesting data set that may potentially be used to answer this question.

Acknowledgements

We are grateful to the publisher Spotlight Verlag GmbH for making their publications available to us for research purposes.

References

Kepa Bengoetxea, Itziar González-Dios, and Amaia Aguirregoitia. 2020. AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural*, 64:61–68.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

Marc Benzahra and Yvon François. 2019. Measuring text readability with machine comprehension: a pilot study. In *Workshop on Building Educational Applications Using NLP*, pages 412–422.

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2018. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518–543.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second language research*, 35(1):99–119.

Tim Vor der Brück and Sven Hartrumpf. 2007. A semantically oriented readability checker for German. In *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznań, Poland. Wydawnictwo Poznańskie.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011a. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58:412–424.

Marc Brysbaert, Emmanuel Keuleers, and Boris New. 2011b. Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2(27).

Xiaobin Chen. 2018. *Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research*. Ph.D. thesis, Eberhard Karls Universität Tübingen Germany.

Xiaobin Chen and Detmar Meurers. 2017. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Xiaobin Chen and Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer-Assisted Language Learning*, 32(4):418–447. https://doi.org/10.1080/09588221.2018.1527358.

Maria Chinkina, Madeeswaran Kannan, and Detmar Meurers. 2016. Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 7–12, Berlin, Germany. Association for Computational Linguistics. http://anthology.aclweb.org/P16-4002.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of past, present, and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Scott A. Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1):11–28.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2013. Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose. In *Proceedings of Recent Advances in Natural Language Processing*.

Sabrina Dittrich, Zarah Weiss, Hannes Schröter, and Detmar Meurers. 2019. Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 41–56, Turku, Finland.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Nick Ellis and Laura Collins. 2009. Input and second language acquisition: The roles of frequency, form, and function. Introduction to the special issue. *The Modern Language Journal*, 93(3):329–335.

B. Janghorban Esfahani, A. Faron, K. S. Roth, P. P. Grimminger, and J. C. Luers. 2016. Systematic readability analysis of medical texts on websites of German university clinics for general and abdominal surgery. *Zentralblatt fur Chirurgie*, 141(6):639–644.

Rudolf Franz Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. https://www.aclweb.org/anthology/D12-1043.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? Assessing the readability of Basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Petronella Grootens-Wiegers, Martine C. De Vries, Tessa E. Vossen, and Jos M. Van den Broek. 2015.

Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, 37(1):89–117.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Mark A Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbai, India. http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Stephen D. Krashen. 1981. The fundamental pedagogical principle in second language teaching. *Studia Linguistica*, 35(1–2):50–70.

Christoph Kühberger, Christoph Bramann, Zarah Weiss, and Detmar Meurers. 2019. Task complexity in history textbooks. a multidisciplinary case study on triangulation in history education research. *History Education International Research Journal (HEIRJ)*, 16(1). Special Issue on Mixed Methods and Triangulation in History Education Research.

Max Kuhn. 2020. caret: Classification and regression training. R package version 6.0-86.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

James P Lantolf, Stephen L Thorne, and Matthew E Poehner. 2015. Sociocultural theory and second language development. In *Theories in second language acquisition: An introduction*. Routledge New York, NY.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020a. An analysis of transfer learning methods for multilingual readability assessment. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–100.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020b. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. http://aclweb.org/anthology/P/P14/P14-5010.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*.

Philip M. McCarthy and Scott Jarvis. 2007. A theoretical and empirical evaluation of vocd. *Language Testing*, 24:459–488.

Danielle A. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, Cambridge, M.A.

Eleni Miltsakaki and Audrey Troutt. 2007. Readx: Automatic evaluation of reading difficulty of web text. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*, pages 7280–7286, Quebec City, Canada. AACE. http://www.editlib.org/p/26932.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic assessment of language learning materials based on linguistic complexity. In *Proceedings of CICLING 2015- Research in Computing Science Journal Issue (to appear)*. https://arxiv.org/abs/1603.08868.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert Reynolds. 2016. *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications*. Ph.D. thesis, UiT - The Arctic University of Norway.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Ivana Lučić. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. http://aclweb.org/anthology/W12-2019.pdf.

Zarah Weiss, Sabrina Dittrich, and Detmar Meurers. 2018. A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Zarah Weiss and Detmar Meurers. 2019a. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019b. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Zarah Weiss and Detmar Meurers. 2021. Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, 7(1):84–131.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers. 2019. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Honolulu.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 49–57.

Appendix A: List of Selected Features

A.1: Features selected for Spotlight-EN

LEN Number Of Letters, SD Token Length in Letters, Percentage of Word Types with More Than 2 Syllables Length Measures, Number of Word Types with More Than 2 Syllables, SD Sentence Length in Tokens, SD Sentence Length in Syllables, Mean Sentence Length in Syllables

SYN Syntactic Complexity Feature: Dependent clauses per T-unit Clausal, Syntactic Complexity Feature: Mean Length of Noun Phrase Phrasal, SD Local Edit Distance for tokens, SD Global Edit Distance for Lemmas

LEX POS Density Feature: Particle, POS Density Feature: Adjective, POS Density Feature: Past Participle Verb, POS Density Feature: Article, POS Density Feature: Coordinating Conjunction, POS Density Feature: Modifier, POS Proper Noun Density, Corrected TTR, Corrected TTR Adjectives, Suqared TTR Words, Uber index (10) Adjectives, Lexical Verb Variation

MOR MCI-5 for Verbs (5 partitions no repetition), MCI-5 for Nouns (5 partitions no repetition), MCI-10 for Nouns (5 partitions no repetition), MCI-5 for Adjectives (2 partitions with repetition), MCI-5 for Adjectives (2 partitions no repetition), MCI-5 for Nouns (5 partitions with repetition), MCI-5 for Nouns (10 partitions no repetition)

DIS Global Lemma Overlap, Mean Local Noun Overlap (word form-based)

USE Logarithmic Word Frequency (Adj Type), Logarithmic Word Frequency (FW Type),

Logarithmic Word Frequency (SD Adj Token), Logarithmic Word Frequency (SD FW Type), Logarithmic Word Frequency (LW Type), Logarithmic Word Frequency (SD V Type), Logarithmic Word Frequency (AW Type), Word Frequency (AW Type), Logarithmic Word Frequency (V Type), Word Frequency (SD AW Token), Logarithmic Word Frequency (SD LW Token), Word Frequency (FW Token), Logarithmic Word Frequency (SD V Token), Logarithmic Word Frequency (Adv Token), Word Frequency (SD AW Type), Logarithmic Word Frequency (SD LW Type), Word Frequency (SD FW Type)

HLP *none*

A.2: Features selected for Spotlight-DE

LEN Number Of Letters, 2 Number of Word Types with More Than 2 Syllables, Mean Sentence Length in Syllables, SD Sentence Length in Tokens, SD Sentence Length in Letters

SYN Relative Clauses per Sentence, Relative Clauses per Clause, Dependent clauses per Sentence, Dependent clauses per T-unit, Complex T-unit Ratio, Dependent clause ratio, Relative Clauses per T-Unit, Mean Length of T-unit, Verb Cluster per T-Unit, Mean Length of Noun Phrase, Postnominal Modifier per Complex Noun Phrase, Verb Phrases per Clause, Verb Phrases per T-unit

LEX TTR Adverbs per Lexical Types, Squared TTR Nouns, Uber index (10) Verbs, Uber index (10) Nouns, Squared TTR Words, Modals per Verb, POS Modifier Density, POS To-infinitive Density, POS Possessive Pronoun Density, POS Proper Noun Density

MOR MCI-5 for Nouns (2 partitions with repetition), MCI-5 for Nouns (5 partitions with repetition), MCI-10 for Nouns (2 partitions no repetition)

DIS *none*

USE Word Frequency (V Type), Word Frequency (SD V Type), Logarithmic Word Frequency (Adj Token), Logarithmic Word Frequency (SD V Token), Word Frequency (AW Type), Logarithmic Word Frequency (SD Adv Token), Logarithmic Word Frequency (LW Type), Logarithmic Word Frequency (V Token), Word Frequency (SD FW Token), Logarithmic Word Frequency (SD AW Token), Word Frequency (SD AW Type)

HLP *none*

Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?

Zarah Weiss
Department of linguistics
University of Tübingen
Germany
zweiss@sfs.uni-tuebingen.de

Detmar Meurers
Department of linguistics
University of Tübingen
Germany
dm@sfs.uni-tuebingen.de

Abstract

The paper presents a new state-of-the-art sentence-wise readability assessment model for German L2 readers. We build a linguistically broadly informed machine learning model and compare its performance against four commonly used readability formulas. To understand when the linguistic insights used to inform our model make a difference for readability assessment and when simple readability formulas suffice, we compare their performance based on two common automatic readability assessment tasks: predictive regression and sentence pair ranking. We find that leveraging linguistic insights yields top performances across tasks, but that for the identification of simplified sentences also readability formulas – which are easier to compute and more accessible – can be sufficiently precise. Linguistically informed modeling, however, is the only viable option for high quality outcomes in fine-grained prediction tasks.

We then explore the sentence-wise readability profile of leveled texts written for language learners at a beginning, intermediate, and advanced level of German. Our findings highlight that a texts' readability is driven by the maximum rather than the overall readability of sentences. This has direct implications for the adaptation of learning materials and showcases the importance of studying readability also below the document level.

1 Introduction

Comprehensible input is key to foster language learning (Swain, 1985), especially when it challenges learners by falling slightly above their individual level of language competence (Vygotsky, 1978; Krashen, 1985). Also in content-matter education, input comprehensibility has been linked to learning success (e.g., O'Reilly and McNamara, 2007). Thus, automatic readability assessment (ARA) is a crucial tool to support education. ARA seeks to align language input with readers' comprehension skills (Vajjala, 2021; Collins-Thompson, 2014). It can not only identify suitable reading materials, but can also ensure learner-input alignment in applications such as tutoring systems or educational conversational agents or as a validation tool for publishers of educational materials. Yet, most work on ARA focuses on English native speakers, leaving much potential for other languages and approaches specifically tailored to the needs of second or foreign language (L2) learners who experience language barriers differently than native speakers (Greenfield, 2004; Collins-Thompson, 2014).

Although most work on ARA has focused on estimating the readability of entire documents, there are many application scenarios in which sentence-level readability assessment is more suitable. Beyond the identification of text simplification targets (Vajjala and Meurers, 2014), they are also more suitable for very short text types including social media language (e.g., tweets and chats), questionnaire or test items used in assessment and empirical education research, or shorter text units in traditional learning materials (e.g., captions or tasks in schoolbooks). Furthermore, there has been little research on the link between sentence and document readability (but see Vajjala and Meurers, 2014) which is immediately relevant for the targeted design and adaptation of educational materials.

There is a startling gap between the methods proposed in ARA research and those used in practice. While for the last two decades, research on ARA has favored machine learning approaches over traditional readability formulas (Vajjala, 2021) due to their generally better performance (e.g., François and Miltsakaki, 2012), simple formulas continue to be used extensively in practice due to their ease of use and low computation demands (Benjamin, 2012). This discrepancy raises the practical question when simple approximations of readability through formulas suffice, and when the use of more

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 141 - 153

elaborate systems is necessary.

This paper addresses these issues with four major contributions: First, we present a new state-of-the-art (SOTA) sentence-level readability model for L2 German readers which is based on broad linguistic complexity assessment. Its performance on a 7-point Likert scale is comparable to human raters when it comes to estimating the readability of sentences for German L2 readers. Second, we make this model accessible online to enhance the impact of our work outside academic discourse. Users can extract features from their texts using the publicly available web platform CTAP (Chen and Meurers, 2016; Weiss et al., 2021) and use the results as input for a pre-written R script that applies the model to users' input files in the format that is returned by CTAP.[1] Third, we compare our SOTA machine learning-based approach with commonly used readability formulas for the two common ARA tasks predictive regression and ranking to answer the question when using linguistic insights indeed makes a difference and for which tasks simple readability formulas suffice. Finally, we leverage our SOTA model to explore sentence profiles of leveled L2 articles to provide new insights into the role of sentence readability for document difficulty that can help inform input adaptation strategies for educational materials.

The remainder of this paper is structured as follows: after a brief literature review (Section 2), we introduce the data (Section 3) and linguistic features (Section 4) used for our studies. We then report on the model training and evaluation for predictive regression and sentence ranking (Section 5). Finally, we explore the readability profile of German L2 articles on a document level (Section 6) and discuss our overall findings (Section 7). We conclude with finals remarks on the impact of our findings and an outlook on future work (Section 8).

2 Related work

Early approaches to ARA date back to the last century when traditional readability formulas (e.g., Flesch, 1948; Dale and Chall, 1948) were developed, see DuBay (2004, 2006) for a comprehensive overview. Readability formulas estimate text readability solely based on surface level proxies of text characteristics (e.g., sentence and word length or word frequency). They have been heavily criticized for their lack of linguistic insight and robustness, and have been shown to yield inferior results to statistical approaches to ARA on authentic data (François and Miltsakaki, 2012; Collins-Thompson, 2014; Benjamin, 2012; Vajjala, 2021). Yet, they are still the most widely distributed form of ARA in practice due to their low computational demands, ease of use, and availability for a variety of languages (Benjamin, 2012). Common use cases include work on health literacy (Kiwanuka et al., 2017; Grootens-Wiegers et al., 2015; Esfahani et al., 2016) and as evaluation metrics in computational linguistic work on machine translation (Agrawal and Carpuat, 2019; Marchisio et al., 2019; Stymne et al., 2013) or conversational agents (Langevin et al., 2021; Gnewuch et al., 2018; Santhanam et al., 2020).

Since the early 2000s (cf. Vajjala, 2021), statistical approaches became dominant in research on ARA. This includes feature-based approaches leveraging rich linguistic information for their predictions as well as neural approaches without prior feature engineering. While either method has been shown to yield SOTA performances (e.g., Vajjala and Lučić, 2018; Weiss et al., 2021; Martinc et al., 2021; Bengoetxea et al., 2020) on the OneStopEnglish corpus by Vajjala and Lučić (2018), neural approaches have been argued to be more easily applicable for cross-linguistic readability assessment (Martinc et al., 2021; Madrazo Azpiazu and Pera, 2019), but see Weiss et al. (2021); De Clercq and Hoste (2016). Feature-based approaches, instead, are more appropriate when little data is available or when users need an explanation for the obtained readability score, as is commonly the case in education contexts and for publishers of leveled reading materials who might want to revise their texts after obtaining a readability score (Collins-Thompson, 2014). Established features measure aspects of syntax and lexicon (Collins-Thompson, 2014), morphology (Gonzalez-Dios et al., 2014; Hancke et al., 2012; Weiss et al., 2021), and discourse features. They intersect with common features from automatic writing quality assessment (Crossley, 2020) and Second Language Acquisition research (Vajjala and Meurers, 2012).

Only limited progress has been made on ARA for German, after early work on readability formulas (e.g., Amstad, 1978; Björnsson, 1983; Bamberger and Vanecek, 1984). The now unavailable

[1] Both, the complexity-based model and the R script can be accessed at https://osf.io/jg6kc/?view_only=2d62778d592642a4a210eb4c7cc61f87

DeLite system has been used to predict readability for German municipal texts (Vor der Brück and Hartrumpf, 2007; Vor der Brück et al., 2008a,b). Hancke et al. (2012) and Weiss and Meurers (2018) focused on the binary distinction of texts for adult versus young native speaking readers. However, binary ARA is of limited use in practice. Weiss et al. (2021) present to our knowledge the first and only multi-level classification approach for German documents after introducing the first multi-level readability corpus for German, which is part of a larger multi-lingual readability corpus for language learners. For sentence-wise readability assessment, Naderi et al. (2019a) compiled a German corpus of rated sentences and sentence simplification pairs. Naderi et al. (2019b) used this corpus to train a feature-based regression model yielding a root mean squared error (RMSE) of 0.847 which is to our knowledge the current SOTA on this data.

Little research has investigated the relationship between sentence and document readability, even though there has been some work testing the reliability of readability assessment for very short texts (Collins-Thompson and Callan, 2004) and sentences (Dell'Orletta et al., 2011; Vajjala and Meurers, 2014; Pilán et al., 2014). Vajjala and Meurers (2014) inspect readability differences between sentences from Wikipedia and Simple Wikipedia to investigate the poor performance of document-level ARA models for the identification of sentences from simple and regular texts. They find that sentences from Wikipedia are not systematically more complex than sentences from Simple Wikipedia. This raises several questions for further inquiry. The lack of observable differences might be caused by an insufficient sensitivity of the document-level model for sentence-level readability differences. Also, Simple Wikipedia has criticized as not systematically simpler than Wikipedia (e.g., Štajner et al., 2012; Xu et al., 2015; Yaneva et al., 2016). More research is needed to confirm or refute their finding that harder texts are not simply characterized by containing generally less readable sentences which would have direct implications for work on targeted document adaptation seeking to identify language barriers in educational materials.

3 Data

3.1 TextComplexityDE

The TextComplexityDE corpus (Naderi et al., 2019a) consists of 1,119 sentences. 1,019 sentences were extracted from 23 Wikipedia articles related to history, society, or science and 100 sentences from two articles in *Leichte Sprache* (engl. "simple language"). All were rated by 267 German L2 learners along three separate dimensions defined by Naderi et al. (2019a): readability, understandability, and lexical difficulty. For each dimension, sentences were rated by up to ten learners on a 7-point Likert scale. These ratings were aggregated into a single Mean Opinion Score (MOS). For this article, we focus on sentences' readability score (MOS-R).

	Mean	Std.	Min.	Max.
MOS-R	3.02	1.18	1.00	6.33
Words / sent.	20.08	10.62	4.00	63.00
Syll. / word	2.07	0.35	0.96	4.00

Table 1: Summary statistics for the TextComplexityDE sentences including number of words per sentence (sent.), number of syllables (syll.) per word, and the Mean Opinion Score for readability (MOS-R)

Table 1 contains summary statistics for the number of words per sentence sentence, the number of syllables per word, and MOS-R. It shows that MOS-R not quite uses the full range of the scale and that sentences are on average quite long (around 20 words) whereas words are relatively short (around two syllables). Sentence length has a strong Spearman rank correlation with MOS-R score ($r_s = 0.70; p < 0.01$). Word length only exhibits a weak correlation with MOS-R ($r_s = 0.26; p < 0.01$). The current SOTA performance for a ARA model lies at RMSE $= 0.847$ (Naderi et al., 2019b).

Sentence simplification pairs The corpus contains 250 sentence pairs of sentences with MOS-R > 4 sampled from all 23 Wikipedia articles and their simplifications. The texts were manually simplified by 75 native speakers and contain additional meta information on whether the simplification is only slightly or considerably simpler than the original. One sentence could not be successfully simplified and was excluded by us, resulting in 249 sentence pairs with valid simplifications.

3.2 Spotlight-DE

The Spotlight-DE corpus (Weiss et al., 2021) consists of 1.447 leveled articles by the Spotlight publisher. Articles' topics are connected to German politics, culture, and every-day life. The texts tar-

get L2 learners at a beginning ($N = 763$), medium ($N = 509$), or advanced ($N = 175$) level. The publisher aligns these three levels with the levels A2, B1/B2, and C1 of the Common European Framework of Reference (Council of Europe).

The reading levels in this corpus are assigned at the document level rather than at the sentence level. To obtain sentence-wise estimates, we split each article into individual sentences. Table 2 characterizes the resulting sentence-wise corpus. Compared

	Mean	Std.	Min.	Max.
Easy ($n = 16,694$)				
Words / sent.	11.00	5.09	1.00	73.00
Syll. / word	1.71	0.35	0.50	5.00
Medium ($n = 27,522$)				
Words / sent.	12.50	6.26	1.00	60.00
Syll. / word	1.73	0.35	0.33	6.00
Advanced ($n = 11,952$)				
Words / sent.	13.30	6.99	1.00	63.00
Syll. / word	1.78	0.37	0.50	5.50

Table 2: Summary statistics for the Spotlight-DE sentences across document reading levels (easy, medium, advanced) including number of number words per sentence (sent.), number of syllables (syll.) per word

to the TextComplexityDE corpus, sentences are much shorter. Also, there are no systematic differences in either sentence or work length across reading levels and no meaningful Spearman rank correlation between sentence length and article reading level ($r_s = 0.12; p < 0.001$) or word length and article reading level ($r_s = 0.06; p < 0.001$). Thus, unlike many other learner corpora, the SpotlightDE corpus does not rely on surface level simplifications to differentiate between proficiency levels.

4 Feature extraction and selection

We extracted 543 features of linguistic complexity from the linguistic domains of syntax, lexicon, and morphology as well as psycho-linguistic features of text cohesion, language use, and human language processing and surface level text features inspired by traditional readability formulas. All features have a long standing tradition in ARA research (Collins-Thompson, 2014) or in related work on automatic text scoring (Crossley, 2020) and Second Language Acquisition complexity research (Wolfe-Quintero et al., 1998; Housen et al., 2012).

For feature extraction, we used the CTAP system (Chen and Meurers, 2016, `http://ctapweb.com`) which has been extended to facilitate the analysis of German by Weiss et al. (2021). We chose this system, because it is to our knowledge the most extensive available analysis system for German. The underlying feature extraction engine for German has proven highly successful and robust in a variety of education-related tasks including readability assessment (Weiss and Meurers, 2018; Weiss et al., 2021; Kühberger et al., 2019) and work linked to writing quality assessment (Weiss and Meurers, 2019a,b; Weiss et al., 2019; Bertram et al., 2021; Riemenschneider et al., 2021). Also, using a publicly available web-based system increases the re-usability of any model using these features in practice.

4.1 Feature description

The German pipeline used in CTAP is described in detail in Weiss et al. (2021) and Weiss and Meurers (2021). The latter also contains a comprehensive definition of all complexity measures. We will limit ourselves here to summarize the types of features used to represent the individual linguistic domains.

Syntax The system measures 75 syntactic features which can be further distinguished into measures of clausal elaboration (e.g., *dependent clauses per clause* or *sentence coordination ratio*) and measures of phrasal elaboration (e.g., *prenominal modifiers per noun phrase* or *mean length of prepositional phrases*), as well as measures of syntactic variance (e.g., *edit distances between constituency parses* or *coverage of nominal modifier types*). This set also includes measures of specific grammatical patterns that have been associated with comprehension difficulties for non-native speakers of German (e.g., *the percentage of non-subject prefields* which Ballestracci (2010) identified as language barriers for Italian learners of German) and raw counts of syntactic patterns, such as the number of dependent clauses.

Lexicon There are 146 features of lexical complexity which can be further divided into measures of lexical richness (e.g., *MTLD* by McCarthy (2005) as well as different mathematical transformations of the type-token ratio), measures of lexical variation (e.g., *verb variation*), and lexical density (e.g., *noun type-token ratio* and other parts-of-speech specific type-token ratios). This group also

contains also features measuring the overall occurrence of different parts-of-speech such as nouns, verbs, or puncuation marks.

Morphology CTAP measures 64 measures of morphological complexity for German. We extract features of nominal and verbal inflection (e.g., *genitive case per noun*), derivation (e.g., *derived nouns per noun*), and compounding (e.g., *average compound depth*). We also measure the variability of morphological exponents using different parametrizations of the Morphological Complexity Index (MCI; Brezina and Pallotti, 2019).

Cohesion We extract 46 measures of text cohesion and discourse for German. The features used here include explicit measures of cohesion (e.g., *causal connectives per sentence*) as well as implicit measures of cohesion linked to the use of pronouns and repetitions of subjects, objects, or nouns.

Language use The system offers 172 lexical language use features based on external German data bases. CTAP calculates average word frequencies and their standard deviations with and without log transformations and binned in log frequency bands for four frequency data bases that represent different types of language use: frequencies based on the Subtlex-DE data base consisting of movie and TV captions and Google Books 2000 (both Brysbaert et al., 2011), dlexDB frequencies (Heister et al., 2011) based on German newspaper articles, and frequencies and age of active use measures extracted from the Karlsruhe Children's Text corpus (Lavalley et al., 2015) consisting of essays written by German children in first to eighth grade.

Human sentence processing There are 21 measures of human processing that can be calculated for German. Weiss and Meurers (2018) and Weiss et al. (2021) have used features based on the Dependency Locality Theory (DLT; Gibson, 2000) for German readability classification using different weight configurations by Shain et al. (2016).

Surface length We extract 18 surface length features for German that solely rely on the identification of sentences, words, letters, and syllables. These features include the raw number of these constructs as well as means and standard deviations for sentence and word length based on these units, e.g., *mean sentence length in syllables*.

4.2 Feature selection

After extracting these features from the TextComplexityDE corpus, we removed all features with near-zero variance, i.e., all features for which at least 80% of the data exhibit the same value. This is the case for 31.3% of features ($N = 170$) due to near-exclusively zero values (i.e., not occurring in most data). This leaves 373 features for the analysis coming from all feature domains which were used for model training in Study 1 (Section 5).

This considerable reduction in the number of features is to be expected for data that is as short as the sentences in the TextComplexityDE corpus (e.g, Weiss and Meurers (2021) also report a reduction of 50% of complexity features for short texts). For example, only 7 of the 46 cohesion measures are sufficiently variable on this data, because most cohesion measures are calculated across sentence boundaries. Similarly, only 19 of 64 measures of morphological complexity are sufficiently variable, because there is not enough language material to produce a variety of inflectional properties. Conversely, nearly all language use and lexical features as well as most features of phrasal elaboration remain included in the reduced feature set.

5 Sentence-wise readability assessment

5.1 Set-up

We trained and compared several machine learning algorithms[2] using 10-folds cross-validation (10 CV) and the z-transformations of the 373 features selected in Section 4.2. We selected these algorithms based on their use in previous research or their robustness against large feature sets with multi-colinearity. The Bayesian Ridge Regression outperformed the other models and will be discussed in more detail in the following. To evaluate this complexity-based model's (henceforth: CBM) overall performance, we calculated its RMSE and Spearman rank correlation (r_s) during 10 CV (Section 5.2) and compared it against the current SOTA performance on the data (RMSE $= 0.847$, Naderi et al., 2019b). We also used the model to rank the pairs of regular and simplified sentences in TextComplexityDE (Section 5.3). We report the ranking accuracy in terms of the percentage of correctly ranked pairs for all i) pairs irrespec-

[2]Multiple linear regression with backward feature selection, linear support vector machine regression, random forests, Bayesian ridge regression (model averaged), Bayesian generalized linear model, quantile regression with LASSO penalty

tive of their degree of simplification ($N = 249$), ii) weakly simplified pairs ($N = 114$), and iii) strongly simplified pairs ($N = 135$).

In both evaluation steps, we compared the CBM's performance against five alternative models. We trained a Bayesian Ridge Regression model using only surface length measures as predictors as a baseline (henceforth: length-based model or LBM). We additionally use the following widely used readability formulas for both tasks:[3]

- the *Amstad Readability Index* (ARI; Amstad, 1978) which adapts the Flesch Reading Ease (Flesch, 1948) to German native speakers;
- the *Erste Wiener Sachtextformel* (WSF; Bamberger and Vanecek, 1984) designed for expository texts for German native speakers;
- The *LIX readability index* (Björnsson, 1983) which has been designed to align texts with adult native speakers' reading skills across a variety of languages including German; and
- the *Miyazaki EFL Readability Index* (MER; Greenfield, 1999, 2004) which was designed for English L2 readers.[4]

We calculated all formulas using a publicly available python-based readability calculator which we adjusted to use stanza (Qi et al., 2020) instead of NLTK (Bird and Loper, 2004) for segmentation.[5]

5.2 Results for regression with 10 CV

Table 3 shows the RMSE and Spearman rank correlation of the estimates with MOS-R in the TextComplexityDE data. Both, LBM and CBM outperform the current SOTA on the TextComplexityDE data ($RMSE = 0.847$; Naderi et al., 2019b). Our linguistically more informed CBM clearly outperforms the LBM in terms of both, RMSE and correlation. Due to the differences in the predicted scales, we cannot compute the RMSE for the four readability formulas, but the correlation shows that both, the CBM and LBM outperform the formulas.

	CBM	LBM	WSF	LIX	ARI	MER
RMSE	.685	.739	*n.a.*	*n.a.*	*n.a.*	*n.a.*
r_s	.806	.785	.681	.679	-.532	-.666

Table 3: RMSE and Spearman rank correlation between MOS-R and the predictions by CBM, LBM, and the readability formulas.

The correlation of ARI with MOS-R is much lower than for the other formulas. This is unexpected, because all formulas use only sentence and word length features. However, ARI assigns a much larger weight to word length than the other formulas which in turn correlates only weakly with MOS-R in TextComplexity-DE (see Section 3.1).

CBM's prediction error lies at RMSE $= 0.685$ points on the Likert scale. This is comparable to the variance between raters in the TextComplexityDE data. Averaged across all rated sentences the across-rater standard deviation for MOS-R is at $1.03 \pm 0.51; IQR = [0.71; 1.41]$. This shows that the error of our CBM lies even below the acceptable range of disagreement exhibited by human raters.

	CBM	LBM	WSF	LIX	ARI	MER
Acc.	96.0	93.0	93.6	93.6	95.6	96.8
$-$	95.6	92.1	91.1	91.1	95.6	96.5
$+$	96.5	94.1	96.5	96.5	95.6	97.0

Table 4: Overall ranking accuracy (Acc.), ranking accuracy for weakly simplified pairs ($-$), and ranking accuracy for strongly simplified pairs ($+$)

5.3 Results for ranking of sentence pairs

Table 4 shows the results of the sentence ranking experiment. The ranking accuracy for all ARA models lies above 90%. With an overall accuracy of 96%, CBM again outperforms LBM and the readability formulas WSF and LIX. However, ARI and MER perform comparably to CBM despite their weak performance on the previous regression experiment. It seems that word length (which is weighted higher for these two formulas than for the rest) is more informative than sentence length for distinguishing simplified and regular sentences.

To also estimate if the models reflect the degrees of simplification in the data (weak vs. strong), we compare the difference in the predicted readability score between each sentence and its simplified counterpart. The difference should be systematically larger for strongly than for weakly simplified sentences. We test this assumption using significance testing[6] ($\alpha < 0.05$) and by estimating

[3] All formula equations are defined in Appendix A.

[4] We added this formula to include an estimate tailored to L2 readers despite the lack of German L2 readability formulas.

[5] https://github.com/zweiss/RC_Readability_Calculator

[6] We used a two-sided t-test or Wilcoxon Rank Sum and Signed Rank Tests depending on the normality of predictions determined with a Shapiro-Wilk Normality Test ($\alpha < .05$).

the effect size with Cohen's d.[7] We see a significant, small effect for CBM ($p = 0.02; d = 0.31$), LBM ($p = 0.04; d = 0.25$), MER ($p < 0.01; d = -0.36$), ARI ($p < 0.01; d = -0.30$), LIX ($p = 0.02; d = 0.35$), and WSF ($p = 0.01; d = 0.35$), see Appendix B for a visualization of the findings.

6 Exploring text profiles in leveled articles

6.1 Set-up

We used CBM to explore the text profiles of easy, medium, and advanced articles in the Spotlight-DE corpus, because it was the most precise model in Study 1. With CTAP, we extracted the 373 features from the sentence-split Spotlight-DE data that are used by the model and calculated their z-scores. We inspected the distribution of sentence readability scores across article levels from several perspectives. We first compared the overall differences in sentence complexity per article level and the differences in maximum sentence complexity using significance testing, effect size estimation (parallel to Study 1) and data visualization. We then evaluated the proportions of sentences within a 0.5 point sentence readability interval across article levels. Finally, we visualized the sentence readability of the first ten sentences in a sample of Spotlight-DE articles in three heatmaps, one for each article levels annotated in the Spotlight-DE corpus. This way, we obtain a non-aggregated estimate of the text profiles. To keep the heatmaps comparable, we used all 175 advanced articles as well as a random sample of 175 easy and 175 medium articles containing at least ten sentences.

6.2 Results

Figure 1 combines different perspectives on the sentence-wise article profiles split by article level. We see that the prediction ranges from 1 to 5, a reasonable coverage of the empirically observed MOS-R scale ($1 - 6.33$) in the TextComplexityDE data given the corpus characteristics discussed in Section 3. Figure 1a summarizes the overall sentence readability grouped by article levels with notches indicating the 95% confidence interval. There are small significant differences between easy and medium ($p < 0.001; d = -0.259$) and easy and advanced ($p < 0.001; d = -0.435$) articles, but only negligible albeit significant differences medium and advanced ($p < 0.001; d = -0.178$) articles. The boxplot shows considerable overlap for the 50% range of the data even between easy and advanced sentences. In Figure 1b, which considers only articles' maximum sentence readability scores, this overlap is considerably reduced. Here, we observe large significant differences between easy and advanced ($p < 0.001; d = -2.05$) and medium and advanced ($p < 0.001; d = -1.24$) articles, and moderate significant differences medium and advanced ($p < 0.001; d = -0.689$) articles. This indicates that the maximum sentence readability is more indicative for overall readability level of a text than considering the readability of all its sentences.

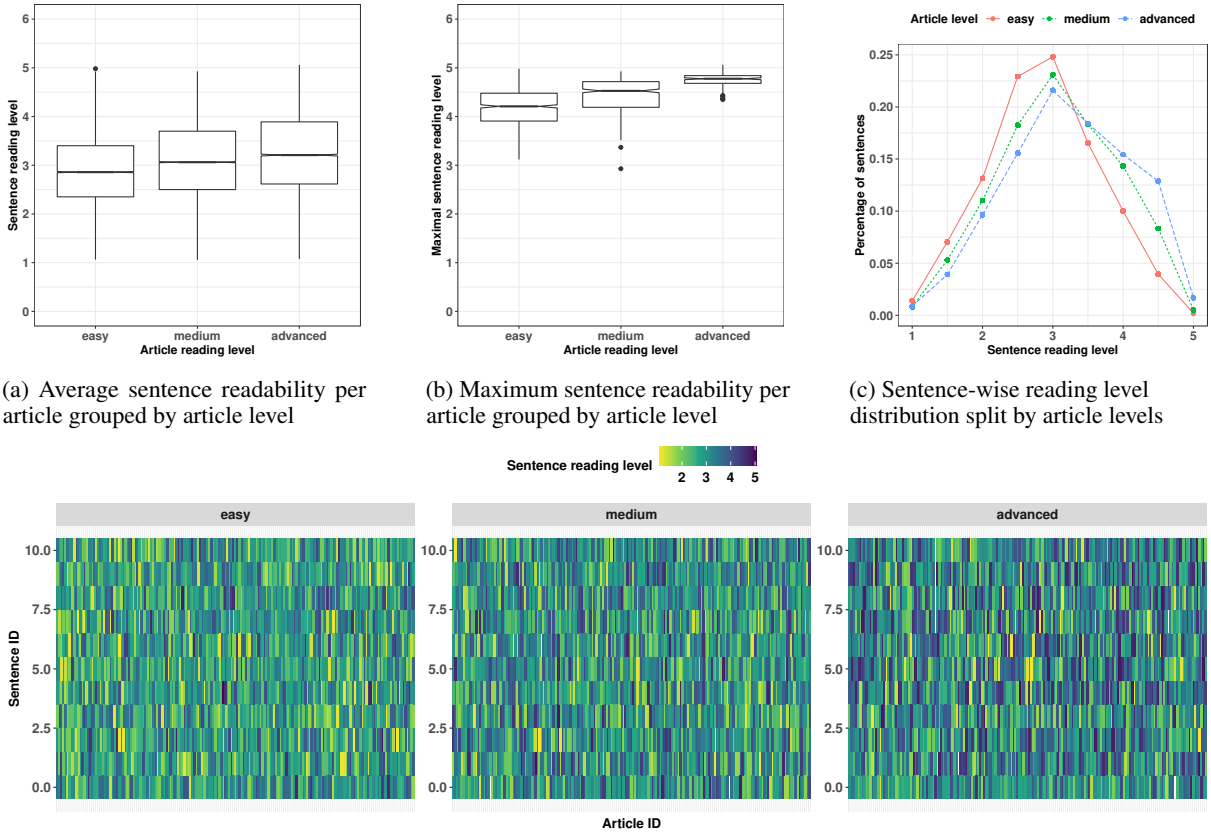
Figure 1c confirms this by comparing the percentage of sentences falling within a 0.5 point readability range across article levels. Sentences from articles at all levels are predominantly medium difficult (MOS-R$= 3$) and between 55.6% (advanced) to 64% (easy) of sentences fall in the range from $2.5 \leq$ MOS-R ≤ 3.5. Article levels differ mostly in the tails of the distribution. The difference is most pronounced for higher difficulty levels (MOS-R ≥ 4): 30% of sentences from advanced articles fall into this range, but only 23.1% of sentences from medium and 14.1% of sentences from easy articles. Even so, it is worth noting that the percentage of sentences with MOS-R ≤ 3 is systematically highest for easy articles and higher for medium than advanced articles. Inversely, the percentage of sentences with MOS-R > 3 is highest for advanced articles and higher for medium than easy articles.

Figure 1d visualizes the sentence readability scores of the first ten sentences of 175 articles per article level. The heatmap depicts the first ten sentences of each sampled article rather than summarizing across sentences and articles at the same article level to demonstrate the relative homogeneity of sentence reading scores for articles at the same article level and the systematic increase in the proportion of more demanding sentences across individual articles with higher article levels.

7 Discussion

Study 1 investigated the performance of linguistically informed readability models and readability formulas for sentence-wise readability assessment for two common ARA tasks: precise predictive regression (Section 5.2) and ranking to identify simplified sentences in sentence simplification pairs (Section 5.3). The results showcase the ver-

[7] We tested for unequal variance using an F test ($\alpha < .05$). In case of unequal variance, we used a Welch approximation for unequal variances to calculate Cohen's d.



(a) Average sentence readability per article grouped by article level

(b) Maximum sentence readability per article grouped by article level

(c) Sentence-wise reading level distribution split by article levels

(d) Predicted sentence readability for the first ten sentences of 175 randomly sampled easy, medium, and advanced articles. Each sentence is represented by a cell. Its readability is encoded with the cell color. The cell's position on the x-axis encodes the article it belongs to and its position on the y-axis its position in that article, e.g., the third sentence in each article is located at $y = 3$.

Figure 1: Sentence readability profiles predicted by our complexity-based model on the Spotlight-DE corpus grouped by article levels (easy, medium, advanced) to showcase differences in sentence readability across documents at different difficulty levels.

satile performance of linguistically informed readability models: only our complexity-based model achieved top performance for both tasks. For the more difficult and authentic task of precise predictive regression, we showed that our linguistically informed complexity-based model clearly outperforms simplistic formulas and set a new SOTA performance (RMSE=0.685) on the data set. The better performance cannot be exclusively attributed to the statistically stronger method, because on both tasks, the complexity-based model clearly outperformed the length-based model. This shows that broad linguistic modeling adds valuable insights beyond the powerful statistical training method.

For ranking, all ARA models achieved an accuracy well above 90% and two readability formulas performed at par with our complexity-based model. This shows that even simple ARA approaches can successfully distinguish relative differences in readability between content-wise equivalent sentences that are being introduced by text simplification. Despite being a rather artificial task, this has some limited applications, e.g., when evaluating machine translation and text simplification systems.

In Study 2, we used our complexity-based model to inspect the sentence-wise readability profiles of leveled texts for L2 readers. Our findings clearly show that while there is a tendency for easier texts to contain more sentence with lower difficulty scores, also medium and advanced texts contain mostly accessible sentences. It is really the presence of difficult sentences within documents that dictates an articles' overall readability. This has clear implications for the design and simplification of educational materials: to efficiently adjust the overall readability level of a text, we need to identify specific sentences that form language barriers rather than simplifying the entire text.

8 Conclusion

We have presented a new SOTA sentence-wise ARA model for German L2 readers which is pub-

licly available and accessible for users with minimal background in R. Leveraging broad linguistic insights, it predicts readability with a margin of error even below the acceptable disagreement range for humans raters. We showed that to flag simplified sentences also traditional readability formulas suffice, but that broad linguistic modeling is needed to obtain the precise predictive readability estimates that are often required in practice (e.g., to adapting learning and teaching materials).

We further explored leveled articles for German L2 readers to illustrate the practical benefits of sentence-level ARA and gain insights into text profiles of leveled documents. Our findings highlight that the readability of texts is driven by the maximum rather than the overall readability of sentences. This has direct implications for the adaptation of teaching materials, which should focus on identifying specific sentences posing language barriers rather than the simplification of all or any sentence in a text. To our knowledge, this is the first time detailed analysis of sentence profiles of leveled reading materials for German. Future work should further explore the implications of this for text simplification, for example using eye-tracking studies. Our work lays the foundation for further research on ARA for German and opens up numerous opportunities for educational applications, such as ARA for captions and task descriptions in school books or the analysis of social media and chat conversations with L2 learners.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

T. Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, University of Zurich.

Sabrina Ballestracci. 2010. Der erwerb von verbzweitsätzen mit subjekt im mittelfeld bei italophonen daf-studierenden. erwerbsphasen, lernschwierigkeiten und didaktische implikationen. *Linguistik online*, 41(1).

Richard Bamberger and Erich Vanecek. 1984. *Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten deutscher Sprache*. Jugend und Volk, Vienna.

Kepa Bengoetxea, Itziar González-Dios, and Amaia Aguirregoitia. 2020. AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural*, 64:61–68.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

Christiane Bertram, Zarah Weiss, Lisa Zachrich, and Ramon Ziai. 2021. Artificial intelligence in history education. linguistic content and complexity analyses of student writings in the cahist project (computational assessment of historical thinking). *Computers and Education: Artificial Intelligence*, page 100038.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL demonstration session*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Carl-Hugo Björnsson. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, pages 480–497.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second language research*, 35(1):99–119.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58:412–424.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 113–119, Osaka, Japan. COLING.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of past, present, and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Scott A. Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1):11–28.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information, Costa Mesa, California.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

B. Janghorban Esfahani, A. Faron, K. S. Roth, P. P. Grimminger, and J. C. Luers. 2016. Systematic readability analysis of medical texts on websites of German university clinics for general and abdominal surgery. *Zentralblatt fur Chirurgie*, 141(6):639–644.

Rudolf Franz Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57. Association for Computational Linguistics.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

Ulrich Gnewuch, Stefan Morana, Carl Heckmann, and Alexander Maedche. 2018. Designing conversational agents for energy feedback. In *International Conference on Design Science Research in Information Systems and Technology*, pages 18–33. Springer.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2014. Making biographical data in wikipedia readable: A pattern-based multilingual approach. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 11–20, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Jerry Greenfield. 1999. *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Ph.D. thesis, Temple Univesity.

Jerry Greenfield. 2004. Readability formulas for efl. *JALT Journal*, 26(1):5–24.

Petronella Grootens-Wiegers, Martine C. De Vries, Tessa E. Vossen, and Jos M. Van den Broek. 2015. Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, 37(1):89–117.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbai, India. http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Julian Heister, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken, and Reinhold Kliegl. 2011. dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, 62:10–20.

Alexis Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Elizabeth Kiwanuka, Raman Mehrzad, Adnan Prsic, and Daniel Kwan. 2017. Online patient resources for gender affirmation surgery: an analysis of readability. *Annals of Plastic Surgery*, 79(4):329–333.

Stephen D Krashen. 1985. *The input hypothesis: Issues and implications*. Longman, New York.

Christoph Kühberger, Christoph Bramann, Zarah Weiss, and Detmar Meurers. 2019. Task complexity in history textbooks. a multidisciplinary case study on triangulation in history education research. *History Education International Research Journal (HEIRJ)*, 16(1). Special Issue on Mixed Methods and Triangulation in History Education Research.

Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Rémi Lavalley, Kay Berkling, and Sebastian Stüker. 2015. Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, University of Memphis.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019a. Subjective assessment of text complexity: A dataset for german language. *arXiv preprint arXiv:1904.07733*.

Babak Naderi, Salar Mohtaj, Karan Karan, and Sebastian Möller. 2019b. Automated text readability assessment for german language: a quality of experience approach. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Tenaha O'Reilly and Danielle S McNamara. 2007. The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American educational research journal*, 44(1):161–196.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 174–184, Baltimore, Maryland, USA. ACL.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Anja Riemenschneider, Zarah Weiss, Pauline Schröter, and Detmar Meurers. 2021. Linguistic complexity in teachers' assessment of german essays in high stakes testing. *Assessing Writing*, 50:100561.

Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *In Proceedings of the First Workshop on Natural Language Processing for Improving Textual Accessibility*. European Language Resources Association (ELRA).

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 375–386.

Merrill Swain. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Susan M. Gass and Carolyn G. Madden, editors, *Input in second language acquisition*, pages 235–253. Newbury House, Rowley, MA.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. http://aclweb.org/anthology/W12-2019.pdf.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*, 165(2):142–222.

Tim Vor der Brück and Sven Hartrumpf. 2007. A semantically oriented readability checker for German. In *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznań, Poland. Wydawnictwo Poznańskie.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008a. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429–435.

Tim Vor der Brück, Hermann Helbig, and Johannes Leveling. 2008b. The readability checker delite. Technical Report Technical Report 345-5/2008, Fakultät für Mathematik und Informatik, FernUniversität in Hagen.

Lev S. Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the Joint 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Zarah Weiss and Detmar Meurers. 2019a. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019b. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Zarah Weiss and Detmar Meurers. 2021. Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, 7(1):84–131.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers. 2019. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Honolulu.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Victoria Yaneva, Irina P. Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the 10h International Conference on Language Resources and Evaluation (LREC)*, pages 293–299.

A Definition of readability formulas

Equation 1 shows the general form of all four readability formulas consisting of an intercept (β_0), a weighted sentence length estimate ($\beta_1 \times SL$), and a weighted word length estimate ($\beta_2 \times WL$).

$$y = \beta_0 + \beta_1 \times SL + \beta_2 \times WL \quad (1)$$

Table 5 shows the respective weights ($\beta_0, \beta_1, \beta_2$) and measurement units for sentence length (SL) and word length (WL). Equation 2 specifies the definition of the composite score for word length used in the *Erste Wiener Sachtextformel*.

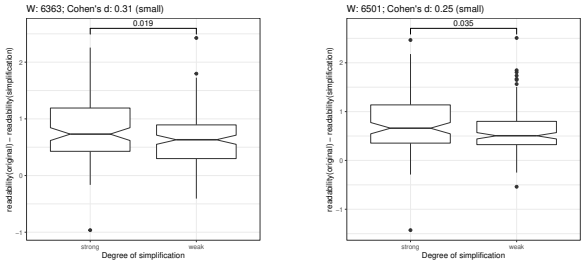
y	β_0	β_1	β_2	SL	WL
LIX	0.0	1.0	1.0	words	syll.
ARI	180.0	−1.0	−58.6	words	syll.
MER	164.9	−1.9	−18.8	words	char.
WSF	0.0	0.2	1.0	words	Eq. 2

Table 5: Weights and measurement units across readability formulas (syll. = syllables, char. = characters)

$$WL_{WSF} = 0.19 \times 3SW + 0.13 \times 6CW - 0.03 \times 1SW - 0.88, \quad (2)$$

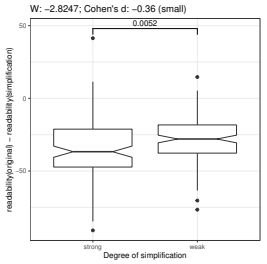
with $3SW$ being the percentage of three or more syllable words, $6CW$ being the percentage of six or more character words, and $1SW$ being the percentage of monosyllabic words. All weights in Table 5 and Equation 1 have been rounded to one decimal point for simplicity.

B Prediction differences between different degrees of simplification

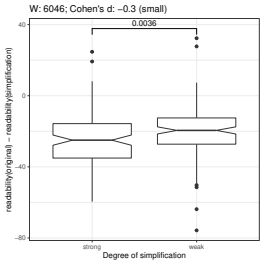


(a) CBM

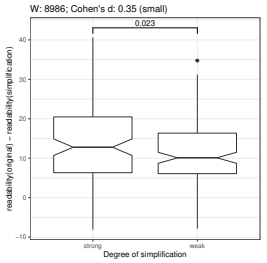
(b) LBM



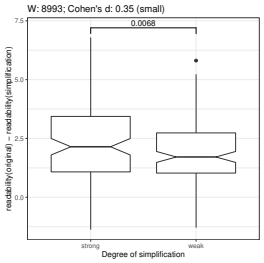
(c) MER



(d) ARI



(e) LIX



(f) EWS

Figure 2: Predicted readability difference between regular and simplified sentences by degree of simplification

Appendix A

Definition of linguistic units

This section defines the central linguistic units that are used for the calculation of complexity measures. A subset of these definitions has also been part of the appendix of Weiss and Meurers (2021), which is also part of this thesis. They are repeated here for easier reference.

Clauses The maximal phrasal projection of a finite verb as well as elliptical constructions that have a sentence-equivalent status.

Complex t-units A t-unit that includes at least one subordinate clause.

Conjunctional clauses Dependent clauses that are introduced by a subordinating conjunction.

Dependent clauses with/without conjunction Conjunctional, interrogative, and relative clauses. Dependent clauses without conjunction are mostly dependent main clauses.

(Graphematic) sentences Strings containing at least one token that are ended by sentence ending punctuation marks or the end of a text.

Half modals The combination of certain verbs with a *zu* (engl. "to") infinitive that they govern. These verbs are: *haben* (engl. "to have"), *sein* (engl. "to be"), *scheinen* (engl. "to seem"), *drohen* (engl. "to threaten"), and *versprechen* (engl. "to promise"). For a more detailed discussion, see Weiss (2015).

Lexical words Nouns, adjectives, adverbs, foreign words, numbers, main verbs, and modal verbs. For a more detailed discussion, see Weiss (2015).

Quasi passives Certain verbs can govern a past participle and result in a construction that serves a similar function as the passive. These words are: *bekommen* (engl. "to get"), *erhalten* (engl. "to receive"), and *kriegen* (engl. "to get"). For a more detailed discussion, see Weiss (2015).

T-units A main clause and all of its dependent clauses and embedded clausal structures (Hunt, 1970, p. 199).

Tokens vs. words I distinguish between tokens and word tokens. Tokens are any continuous string within a text that is separated by white space. This is a simplifying assumption we make for written language. A non-graphematic definition of words and tokens is beyond the scope of this thesis. Words are continuous strings separated by white space which may contain non-alphabetic characters but cannot solely consist of punctuation marks, numbers, formulas, or other symbols.

Types Unique token forms

Non-terminal nodes Non-terminal nodes are all nodes in a tree structure that precede the final (terminal) tree nodes.

Finite clause A clause containing a finite verb (see sentence-equivalent infinitives for different clause types).

***zu*-infinitive** The German *to* infinitive which may or may not be a sentence-equivalent infinitive. For a more detailed discussion, see Weiss (2015).

Complex NP An NP with pre- or post-nominal modifiers.

Complex PP A prepositional phrase (PP) containing an NP with pre- or post-nominal modifiers

Sentence-equivalent infinitives An infinitive construction that approximates a dependent clause. These are often introduced by connectives such as *als*, *anstatt*, *außer*, *ohne*, *statt*, *um*. For a more detailed discussion, see Weiss (2015).

Eventive passives The eventive passive in German is formed with *werden* (engl. "to become"). In contrast, the static passive is formed with *sein* (engl. "to be").

Verb cluster A verb cluster is a group of adjacent verb phrases (VPs) that govern each other. They can take arbitrary sizes but typically do not include more than three verbs (see Weiss, 2015). For a more detailed discussion, see Weiss (2015).

Post- and pre-nominal modifiers Prenominal modifiers precede the noun kernel they modify. Examples are attributive participles, possessive attributes, and adjectives. Postnominal modifiers follow the noun kernel they modify. Examples include PPs, relative clauses, and comparative modifiers. For a more detailed discussion, see Weiss (2015).

Mittelfeld A position in the topoligical field model. Located between the left and right sentence brackets. For a more detailed discussion, see Weiss (2015).

Vorfeld A position in the topoligical field model. It precedes the left sentence bracket. For a more detailed discussion, see Weiss (2015).

Deagentivation patterns Deagentivation is a strategy to obtain a non-personal writing style that omits the subject. It is typically used in (German) academic language to suggest objectivity (Hennig and Niemann, 2013; Polenz, 1981). Common deagentivation devices include passivization, the use of infinitives and participle constructions, and—in German—the use of *man* or *sich-lassen*. For a more detailed discussion, see Weiss (2015).

GermaNet synset A set of lexical units that are connected through semantic relations. For a more detailed discussion, see Hancke (2013).

GermaNet relations Semantic relations between lexical units such as part-whole, antonymy, or hyponymy. For a more detailed discussion, see Hancke (2013).

GermaNet frames Subcategorization information for verbs. For a more detailed discussion, see Hancke (2013).

GermaNet lexical units Elements in the semantic network representing a word sense of a lemma. For a more detailed discussion, see Hancke (2013).

Appendix B

Complexity features

This section contains definitions of all complexity measures that are used in this thesis, be it in the legacy system, the CTAP system, or both. Tables are sorted by linguistic complexity domains through subsections. Raw counts are omitted in the tables for bevity. Both systems can output the raw counts used throughout the feature definitions as features. However, non-normalized counts were generally not used throughout the studies reported in this thesis unless specified otherwise in the respective articles. The tables and definitions may refer to additional information provided in other publications to specify how a feature is calculated. Furthermore, all tables are based on the definition of linguistic units in Section A. These definitions are not re-iterated in the tables.

Each table indicates whether or not a feature is included in the legacy system or the German CTAP code. Features that are included in a system are marked with a check mark in the corresponding column (✓) or otherwise marked with a cross (✗). Feature names used throughout articles or systems may slightly vary from the names used here (e.g., `mean sentence length in words` versus `number of words per sentence`). Some features sharing the same nominator or denominator are grouped into a single row for brevity. In these cases, the definition column introduces numbers and letters to reference specific numerators/denominators. The columns 'legacy' and 'CTAP' then specify which of the combinations are available in the respective systems by using the respective number/letter combinations without the check mark.

B.1 Syntactic complexity measures

This section contains definitions of all syntactic complexity measures that are used in this thesis. Features are grouped into tables based on their complexity sub-domains. This section

contains the following tables:

Table B.1 contains all measures of global syntactic elaboration.

Table B.2 contains all measures of clausal syntactic elaboration.

Table B.3 contains all measures of phrasal syntactic elaboration and variation.

Table B.4 contains all additional measures of syntactic complexity that target sub-clausal units that do not focus on individual phrases. This entails predominantly grammatical patterns and periphrastic tenses.

Table B.5 contains all measures of syntactic variation.

Table B.1: *Global syntactic complexity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Mean sentence length in letters	#letters ÷ #sentences	✗	✓
Mean sentence length in syllables	#syllables ÷ #sentences	✗	✓
Mean sentence length in words	#words ÷ #sentences	✓	✓
Longest sentence in words	max(#words in one sentence)	✓	✗
Average number of non-terminal nodes	$\sum$ #non-terminal nodes in a parse tree ÷ X; **with X being: #sentences [1], #t-units [2], #clauses [3], or #finite clauses [4]**	1–4	✗
Average parse tree height	$\sum_i$ maximal number of non-terminal nodes between the root node to a terminal node in the constituency tree of sentence i ÷ X; with options for X being defined above	1–4	✗
Mean length of clause	#words ÷ #clauses	✓	✓
Mean length of finite clause	#words ÷ #finite clauses	✓	✗
Mean length of complex t-unit	#words ÷ #complex t-units	✗	✓
Mean length of t-unit	#words ÷ #t-units	✓	✓

Table B.2: *Clausal syntactic complexity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Complex t-unit ratio	#complex t-units ÷ #t-units	✓	✓
Complex t-unit per sentence	#words ÷ #sentences	✓	✓
Dependent clause ratio	#dependent clauses ÷ #clauses	✓	✓
Dependent clauses per sentence	#dependent clauses ÷ #sentences	✓	✓
Dependent clauses per t-unit	#dependent clauses ÷ #t-units	✓	✓
Dependent clauses per finite clause	#dependent clauses ÷ #finite clauses	✓	✗
Relative clauses per clause	#relative clauses ÷ #clauses	✓	✓
Relative clauses per finite clause	#relative clauses ÷ #finite clauses	✓	✗
Relative clauses per dependent clause with conjunction	#relative clauses ÷ #dependent clauses with conjunction	✓	✗
Relative clauses per sentence	#relative clauses ÷ #sentences	✓	✓
Relative clauses per t-unit	#relative clauses ÷ #t-units	✓	✓
Sentence complexity ratio	#clauses ÷ #sentences	✓	✓
Sentence coordination ratio	#t-units ÷ #sentences	✓	✓
T-unit complexity ratio	#clauses ÷ #t-units	✓	✓
Dependent clause with conjunction ratio	#dependent clauses with conjunction ÷ X; **with X being: #sentences [1], #t-units [2], #clauses [3], #finite clauses [4], or #dependent clauses with conjunction [5]**	1–5	✗
Conjunctional clause ratio	#conjunctional clauses ÷ X; with options for X being defined above	1–5	✗
Dependent clause without conjunction ratio	#clauses without conjunction ÷ X; with options for X being defined above	1–5	✗
Interrogative clause ratio	#interrogative clauses ÷ X; with options for X being defined above	1–5	✗

Table B.3: *Phrasal syntactic complexity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Complex NPs per sentence/ t-unit/ clause/ finite clause	#complex NPs ÷ X; **with X being: #sentences [1], #t-units [2], #clauses [3], #finite clauses [4]; #dependent clauses [5]**	1–4	1–3
Complex Nominals per NP	#complex NPs ÷ #NPs	✗	✓
NPs per sentence/ t-unit/ clause/ finite clause	#NPs ÷ X; with options for X being defined above	1–4	1–3
PPs per sentence/ t-unit/ clause/ finite clause	#PPs ÷ X; with options for X being defined above	1–4	1–3
VPs per sentence/ t-unit/ clause/ finite clause	#VPs ÷ X; with options for X being defined above	1–4	1–3
zu-infinitives per sentence/ t-unit/ clause/ finite clause	#*zu*-infinitives ÷ X; with options for X being defined above	1–4	✗
Sentence-equivalent infinitives per sentence/ t-unit/ clause/ finite clause/ dependent clause	#sentence equivalent infinitives ÷ X; with options for X being defined above	1–5	✗
Eventive passives per sentence/ t-unit/ clause/ finite clause	#eventive passives ÷ X; with options for X being defined above	1–4	✗
Complex PPs per sentence/ t-unit/ clause	#complex NPs ÷ X; with options for X being defined above	✗	1–3
Coordinate Phrases per sentence/ t-unit/ clause/ finite clause	#coordinate phrases ÷ X; with options for X being defined above	1–4	1–3
Mean Length of NP	#words in an NP ÷ #NPs	✓	✓
Mean Length of PP	#words in a PP ÷ #PPs	✓	✓
Mean Length of VP	#words in a VP ÷ #VPs	✓	✗
Verb cluster per clause	#verb cluster ÷ #clauses	✗	✓
Verb cluster per sentence	#verb cluster ÷ #sentences	✗	✓
Mean length of verb cluster	#verbs in a verb cluster ÷ #verb clusters	✓	✓

Table B.3: *Phrasal syntactic complexity features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Postnominal modifier per complex NP	#postnominal modifier ÷ #complex NPs	✓	✓
Prenominal modifier per complex NP	#prenominal modifier ÷ #complex NPs	✓	✓
Attributive participles per NP	#attributive participles ÷ #NPs	✓	✗
Clausal noun modifiers per NP	#clausal noun modifiers ÷ #NPs	✓	✗
Comparative noun modifiers per NP	#comparative noun modifiers ÷ #NPs	✓	✗
Determiners per NP	#determiners ÷ #NPs	✓	✗
Possessive noun modifiers per NP	#possessive noun modifiers ÷ #NPs	✓	✗
Average number of noun modifiers	#noun modifier ÷ #nouns	✓	✗
Average number of verb modifiers	#verb modifier ÷ #verbs	✓	✗
Average number of noun dependents	#noun dependents ÷ #nouns with dependents	✓	✗
Average number of verb dependents	#verb dependents ÷ #verbs with dependents	✓	✗
Average number of verb dependents excluding modal verbs	#verb dependents excluding modal verbs ÷ #verbs with dependents excluding modal verbs	✓	✗
Adjective and adverb verb modifiers per VP	(#adjective modifiers of verbs + #adverb modifiers of verbs) ÷ #VPs	✓	✗
Particle verb modifiers per VP	#particle verb modifers ÷ VPs	✓	✗
Prepositional verb modifiers per VP	#prepositional verb modifers ÷ VPs	✓	✗
Standard deviation of verb cluster sizes	standard deviation corresponding to `Mean length of verb cluster` feature defined above	✓	✗

Table B.3: *Phrasal syntactic complexity features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Percentage of main verb clusters	#verb clusters headed by a main verb ÷ verb clusters	✓	✗
Percentage of modal verb clusters	#verb clusters headed by a modal verb ÷ verb clusters	✓	✗
Percentage of auxiliary verb clusters	#verb clusters headed by an auxiliary verb ÷ verb clusters	✓	✗

Table B.4: *Other sub-clausal syntactic complexity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Percentage of periphrastic tenses used	#periphrastic tenses used at least once (present perfect, past perfect, future 1, future 2) ÷ #finite verbs	✓	✗
Percentage of present perfect used	#present perfect used ÷ #finite verbs	✓	✗
Percentage of past perfect used	#past perfect used ÷ #finite verbs	✓	✗
Percentage of Future I used	#future I used ÷ #finite verbs	✓	✗
Percentage of Future II used	#future II used ÷ #finite verbs	✓	✗
Percentage of simple present used	#simple present used ÷ #finite verbs	✓	✗
Percentage of simple past used	#simple past used ÷ #finite verbs	✓	✗
Average *Mittelfeld* (engl. “middle field”) length in syllables	#syllables in the *Mittelfeld* (engl. “middle field”) ÷ #*Mittelfelder* (engl. “middle fields”)	✓	✗

Table B.4: *Other sub-clausal syntactic complexity features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Syllables between first verb argument and main verb (excluding adjacent arguments)	#syllables between the first verb argument and the main verb when the first argument is not immediately preceding or following the main verb ÷ #main verbs that are not adjacent to their first argument		
Percentage of non-subject *Vorfelder* (engl. "prefields")	#*Vorfelder* (engl. "prefields") that do not contain the subject ÷ #*Vorfelder* (engl. "prefields")	✓	✗
man occurrences per subject	#*man* is used as an impersonal subject ÷ #subjects	✓	✗
Percentage of infinitival constructions	#infinitives ÷ #VPs	✓	✗
sich-lassen occurrences per subject	#*sich-lassen* is used to drop the subject ÷ #VPs	✓	✗
Percentage of half modals	#half modals÷ #VPs	✓	✗
Ratios of passive constructions	#passives ÷ X; **with options for X being: #sentences [1], #t-units [2], #clauses [3], #finite clauses [4], #VPs [5]**	1–5	✗
Ratio of quasi passive constructions	#quasi passives ÷ X; with options for X defined above	1–5	✗
Percentage of deverbal nouns	#deverbal nouns ÷ #NPs	✓	✗

Table B.5: *Syntactic variation features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as shorthand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Mean global/local edit distance	global/local edit distance for X ÷ #parse trees (see Chen, 2018, for details); **with options for X being: lemmas [1], POS [2], tokens [3]**	✗	1–3
Standard deviation of global/local edit distance	standard deviation corresponding to `Mean global/local edit distance` feature defined above	✗	1–3
Standard deviation of sentence length in letters	standard deviation corresponding to `Mean sentence length in letters` defined in Table B.1	✗	✓
Standard deviation of sentence length in syllables	standard deviation corresponding to `Mean sentence length in syllables` defined in Table B.1	✗	✓
Coverage noun modifier types	#noun modifier types occurring at least once (determiners, possessive attributes, prenominal attributes, postnominal attributes, attributive participles, comparative modifiers, clausal modifiers) ÷ #noun modifier types measured ($N = 7$)	✓	✗
Coverage of verb modifier types	$\sum$ verb modifier types occurring at least once (i.e., adjective/adverbial modifiers, PP modifiers, past/present participle, verb particles) ÷ # verb modifier types measured ($N = 4$)	✓	✗
Coverage of verb cluster sizes	$\sum$ verb cluster size occurring at least once (covering sizes 2 to ≥ 6) ÷ # verb cluster sizes measured ($N = 5$)	✓	✗

Table B.5: *Syntactic variation features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Coverage of verb cluster types	$\sum$ verb cluster types occurring at least once (auxiliary, main, modal) $\div$ # verb cluster sizes measured ($N = 3$)	✓	✗
Coverage of tenses	#tenses occurring at least once (present perfect, past perfect, future 1, future 2, simple present, simple past) $\div$ #tenses measured ($N = 6$)	✓	✗
Coverage of periphrastic tenses	#periphrastic tenses occurring at least once (present perfect, past perfect, future 1, future 2) $\div$ #periphrastic tenses measured ($N = 4$)	✓	✗
Coverage of deagentivation patterns	$\sum$ deagentivation patterns occurring at least once (*man* occurrences, infinitival constructions, *sich-lassen* occurrences, half modal clusters, passives, quasi passives, participle modifiers, attributive participles) $\div$ #deagentivation patterns measured ($N = 8$)	✓	✗

B.2 Lexical complexity measures

This section contains definitions of all lexical complexity measures (excluding language use features) that are used in this thesis. Features are grouped into tables based on their complexity sub-domains. This section contains the following tables:

Table B.6 contains all measures of global lexical complexity.

Table B.7 contains all measures of lexical diversity.

Table B.8 contains all measures of lexical density.

Table B.6: *Global lexical complexity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Mean word length in letters	#letters ÷ #word tokens	✓	✓
Mean word length in syllables	#syllables ÷ #word tokens	✓	✓
Percentage of words with more than 2 syllables	#word tokens with more than two syllables ÷ #word tokens	✗	✓
Percentage of word types with more than 2 syllables	#word types with more than two syllables ÷ #word types	✗	✓
Maximal word length in syllables	max(#syllables per word)	✓	✗
Standard deviation of word length in letters	standard deviation corresponding to `Mean word length in letters` defined above	✗	✓
Standard deviation of word length in syllables	standard deviation corresponding to `Mean word length in syllables` defined above	✗	✓

Table B.7: *Lexical diversity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
HD-D (excluding punctuation and numbers)	see McCarthy and Jarvis (2007)	✓	✓
MTLD (excluding punctuation and numbers)	see McCarthy and Jarvis (2010)	✓	✓
Yule's k	$10^4 * \frac{(\sum fX*X)-\#tokens}{\#tokens^2}$, with X being the frequency vector of each word type and fX being the frequency of each type frequency in X	✓	✗
Type Token Ratio (10 Segments)	TTR averaged across 10 segments	✗	✓
Type Token Ratio (50 Segments)	TTR averaged across 50 segments	✗	✓
Corrected TTR	$\sqrt{\frac{\#types}{2*\#tokens}}$ [1] or $\frac{\#types}{\sqrt{2*\#tokens}}$ [2]	1	2
Corrected TTR (excluding punctuation and numbers)	$\frac{\#types}{\sqrt{2*\#tokens}}$ (excluding punctuation and numbers)	✗	2
Corrected TTR for Adjectives/ Adverbs/ Lexical Words/ Nouns/ Verbs/ Words	as `Corrected TTR` [2] but using the respective POS-specific types and tokens	✗	✓
Log TTR	log(#types) ÷ log(#tokens)	✗	✓
Bilogarithmic TTR	$\log_2$(#types) ÷ $\log_2$(#tokens)	✓	✗
Log10 TTR	$\log_{10}$(#types) ÷ $\log_{10}$(#tokens)	✗	✓
Log10 TTR (excluding punctuation and numbers)	$\log_{10}$(#types) ÷ $\log_{10}$(#tokens); excluding punctuation and numbers	✗	✓
Log10 TTR for Adjectives/ Adverbs/ Lexical Words/ Nouns/ Verbs/ Words	as `Log10 TTR` but using the respective POS-specific types and tokens	✗	✓
Root TTR	$\sqrt{(\#types \div \#tokens)}$ [1] or $\#types \div \sqrt{\#tokens}$ [2]	1	2
Root TTR (excluding punctuation and numbers)	$\#types \div \sqrt{\#tokens}$ (excluding punctuation and numbers)	✗	✓

Table B.7: *Lexical diversity features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Root TTR for Adjectives/ Adverbs/ Lexical Words/ Nouns/ Verbs/ Words	as `Root TTR` [2] but using the respective POS-specific types and tokens	✗	✓
Squared TTR (excluding punctuation and numbers)	$\#types^2 \div \#tokens$ (excluding punctuation and numbers)	✗	✓
Squared TTR	$\#types^2 \div \#tokens$	✗	✓
Squared TTR for Adjectives/ Adverbs/ Lexical Words/ Nouns/ Verbs/ Words	as `Squared TTR` but using the respective POS-specific types and tokens	✗	✓
TTR	#types ÷ #tokens	✓	✓
TTR (excluding punctuation and numbers)	#types ÷ #tokens (excluding punctuation and numbers)	✗	✓
TTR for Adjectives/ Adverbs/ Lexical Words/ Nouns/ Verbs/ Words	as `TTR` but using the respective POS-specific types and tokens	✗	✓
Uber index	$log(\#tokens)^2 \div log(\frac{\#types}{\#tokens})$	✓	✓
Uber10	$log_{10}(\#tokens)^2 \div log_{10}(\frac{\#types}{\#tokens})$	✗	✓
Uber10 for Adjectives/ Adverbs/ Lexical Words/ Nouns/ Verbs/ Words	as `Uber10` but using the respective POS-specific types and tokens	✗	✓
Lexical TTR	(#lexical types ÷ #lexical tokens)	✓	✓
Lexical TTR for adjectives and adverbs/ adjectives/ adverbs/ nouns/ verbs	(#lexical types ÷ #lexical tokens)	✓	✓
Lexical variation I	#lexical types ÷ #lexical tokens	✓	✓
Lexical variation II	#lexical types ÷ #tokens	✓	✗
Verb variation I	#verb types (excl. auxiliary verbs) ÷ #verb tokens (excl. auxiliary verbs)	✓	✓

Table B.7: *Lexical diversity features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Corrected verb variation I	#verb types (excl. auxiliary verbs) $\div\sqrt{2 * \#verb\ tokens}$ (excl. auxiliary verbs)	✓	✓
Squared verb variation I	#verb types^2 (excl. auxiliary verbs) ÷ #verb tokens (excl. auxiliary verbs)	✓	✓
Verb variation II	#verb types (excl. auxiliary verbs) ÷ #lexical tokens (excl. auxiliary verbs)	✓	✓
Lexical variation for adjectives/ adverbs/ nouns/ modifiers	X ÷ #lexical tokens; **with options for X being: #adjective types [1], #adverb types [2], #noun types [3], #modifier types [4], #adjective and adverb types [5]**	1–3, 5	1–4
Verbs per token	#verb types (excl. auxiliary verbs) ÷ #tokens	✓	✗
Nouns per token	#noun types ÷ #tokens	✓	✗

Table B.8: *Lexical density features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Lexical words per word	#lexical words ÷ #word tokens	✗	✓
Modals per verb	#modal verbs ÷ #verb tokens	✗	✓
Modals per word	#modal words ÷ #word tokens	✗	✓
Verb to noun ratio	#verb tokens ÷ #noun tokens	✓	✗
haben instances per verb	#*haben* (engl. "to have") tokens ÷ #verb tokens	✓	✗
sein instances per verb	#*sein* (engl. "to be") tokens ÷ #verb tokens	✓	✗
Lexical density of adjectives	#adjective tokens ÷ #word tokens	✗	✓

Table B.8: *Lexical density features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Lexical density of adverbs	#adverb tokens ÷ #word tokens	✗	✓
Lexical density of articles	#article tokens ÷ #word tokens	✗	✓
Lexical density of auxiliary verbs	#auxiliary verb tokens ÷ #word tokens	✗	✓
Auxiliary Verbs per verb	#auxiliary verb tokens ÷ #verb tokens	✓	✗
Modal verbs per verb	#modal verb tokens ÷ #verb tokens	✓	✗
Lexical density of cardinal numbers	#cardinal number tokens ÷ #word tokens	✗	✓
Lexical density of common nouns	#common noun tokens ÷ #word tokens	✗	✓
Lexical density of comparative conjunctions	#comparative conjunction tokens ÷ #word tokens	✗	✓
Lexical density of conjunctions	#conjunction tokens ÷ #word tokens	✗	✓
Lexical density of coordinating conjunctions	#coordinating conjunction tokens ÷ #word tokens	✗	✓
Lexical density of demonstrative pronouns	#demonstrative pronoun tokens ÷ #word tokens	✗	✓
Lexical density of determiners	#determiner tokens ÷ #word tokens	✗	✓
Lexical density of finite verbs	#finite verb tokens ÷ #word tokens	✗	✓
Lexical density of foreign words	#foreign word tokens ÷ #word tokens	✗	✓
Lexical density of functional words	#functional word tokens ÷ #word tokens	✗	✓
Lexical density of indefinite pronouns	#indefinite pronoun tokens ÷ #word tokens	✗	✓
Lexical density of infinite verbs	#infinite verb tokens ÷ #word tokens	✗	✓
Lexical density of interjections	#interjection tokens ÷ #word tokens	✗	✓
Lexical density of interrogative pronouns	#interrogative pronoun tokens ÷ #word tokens	✗	✓
Lexical density of lexical words	#lexical word tokens ÷ #word tokens	✗	✓

Table B.8: *Lexical density features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Lexical density of main verbs	#main verb tokens ÷ #word tokens	✗	✓
Lexical density of modal verbs	#modal verb tokens ÷ #word tokens	✗	✓
Lexical density of modifiers	#modifier tokens ÷ #word tokens	✗	✓
Lexical density of non-finite verbs	#non-finite verb tokens ÷ #word tokens	✗	✓
Lexical density of nouns	#noun tokens ÷ #word tokens	✗	✓
Lexical density of particles	#particle tokens ÷ #word tokens	✗	✓
Lexical density of past participle verbs	#past participle verb tokens ÷ #word tokens	✗	✓
Lexical density of personal pronouns	#personal pronoun tokens ÷ #word tokens	✗	✓
Lexical density of possessive pronouns	#possessive pronoun tokens ÷ #word tokens	✗	✓
Lexical density of prepositions	#preposition tokens ÷ #word tokens	✗	✓
Lexical density of pronouns	#pronoun tokens ÷ #word tokens	✗	✓
Lexical density of proper nouns	#proper noun tokens ÷ #word tokens	✗	✓
Lexical density of punctuation marks	#punctuation tokens ÷ #word tokens	✗	✓
Lexical density of relative pronouns	#relative pronoun tokens ÷ #word tokens	✗	✓
Lexical density of singular proper nouns	#singular proper noun tokens ÷ #word tokens	✗	✓
Lexical density of subordinating conjunctions	#subordinating conjunction tokens ÷ #word tokens	✗	✓
Lexical density of to infinitives	#*zu*-infinitive tokens ÷ #word tokens	✗	✓
Lexical density of verbs	#verb tokens ÷ #word tokens	✗	✓

B.3 Language use

This section contains definitions of all language use measures (i.e., relative lexical complexity) that are used in this thesis. Features are grouped into tables based on their complexity sub-domains. This section contains the following tables:

Table B.9 contains all measures based on word frequencies.

Table B.10 contains all measures based on word familiarity or informativeness.

Table B.11 contains all measures based on age of active use in written language.

Table B.9: *Frequency features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Average frequency in frequency data-base X (word tokens of type Y)	(∑ frequencies of all word tokens in the text that were found in frequency data-base X) ÷ #word tokens in the text found in frequency data-base X; **frequency data-base options for X:** ***dlexDB [1], Google Books 2000 [2], SUBTLEX-DE [3], KCT corpus [4], OpenSubtitles [5]***; **word type options for Y:** ***all words [a], lexical words [b], function words [c], adjectives [d], adverbs[e], nouns [f], verbs [g]***	1a, 2a, 3a, 4a	2a–c
Average frequency in frequency data-base X (all word types of type Y)	Equivalent to `Average frequency in frequency data-base X (word tokens of type Y)`	1a, 2a, 3a, 4a	2a–c
Average frequency in frequency data-base X (all lemma types)	Equivalent to `Average frequency in frequency data-base X (word tokens of type Y)`	1a, 4a	✗

Table B.9: *Frequency features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Annotated frequency of word types in frequency data-base X	(∑ POS-specific frequencies of all word types in the text that were found in frequency data-base X) ÷ #word types in the text found in frequency data-base X; same frequency data base options as defined above	1	✗
Average frequency per million words in frequency data-base X (word tokens of type Y)	(∑ frequencies per million words of all word tokens in the text that were found in frequency data-base X) ÷ #word tokens in the text found in frequency data-base X; **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	3a–g, 5a–g
Average frequency per million words in frequency data-base X (all word types of type Y)	Equivalent to `Average frequency in frequency data-base X (word tokens of type Y)`	✗	3a–g, 5a–g
Standard deviation of the frequency per million words in frequency data-base X (all word tokens of type Y)	SD(frequency per million words of all word tokens in the text that were found in frequency data-base X); **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	3a–g, 5a–c
Standard deviation of the frequency per million words in frequency data-base X (all word types of type Y)	Equivalent to `Standard deviation of the frequency per million words in frequency data-base X (all word tokens of type Y)`	✗	3a–g, 5a–g

Table B.9: *Frequency features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Average log frequency in frequency data-base X (all word tokens of type Y)	(∑ log frequencies of all word tokens in the text that were found in frequency data-base X) ÷ #word tokens in the text found in frequency data-base X; **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	2a–c, 3a–g, 5a–c
Average log frequency in frequency data-base X (all word types of type Y)	Equivalent to `Average log frequency in frequency data-base X (all word tokens of type Y)`	1a, 2a, 3a, 4a	2a–g, 3a–g, 5a–g
Average log lemma frequency in frequency data-base X	Equivalent to `Average log frequency in frequency data-base X (all word tokens of type Y)` but using lemma types	1a, 2a, 3a, 4a	✗
Average annotated log frequency in frequency data-base X (all word types of type Y)	Equivalent to `Average log frequency in frequency data-base X (all word tokens of type Y)` but for POS-specific frequencies	1a	✗
Standard deviation of the log frequency in frequency data-base X (all word tokens of type Y)	SD(log frequencies of all word tokens in the text that were found in frequency data-base X); **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	3a–g, 5a–c

Table B.9: *Frequency features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Standard deviation of the log frequency in frequency data-base X (all word types of type Y)	Equivalent to `Standard deviation of the log frequency in frequency data-base X (all word tokens of type Y)`	✗	3a–g, 5a–g
Word types found in frequency data-base X	#word types found in frequency data-base X ÷ #word types in the text with options for X being defined above	1, 3, 4	✗
Word types not found in frequency data-base X	#word types not found in frequency data-base X ÷ #word types in the text with options for X being defined above	1, 3, 4	✗
Lemma types found in frequency data-base X	#lemma types found in frequency data-base X ÷ #lemma types in the text with options for X being defined above	4	✗
Lexical lemma types found in frequency data-base X	#lexical lemma types found in frequency data-base X ÷ #lexical lemma types in the text with options for X being defined above	4	✗
Average log frequency per million words in frequency data-base X (all word tokens of type Y)	(∑ log frequencies per million words of all word tokens in the text that were found in frequency data-base X) ÷ #word tokens in the text found in frequency data-base X; **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	2a–c, 3a–c

Table B.9: *Frequency features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Average log frequency per million words in frequency data-base X (all word types of type Y)	Equivalent to `Average log frequency per million words in frequency data-base X (all word tokens of type Y)`	✗	2a–c, 3a–c
Log annotated type frequency for frequency band Y based on frequency data-base X	#POS-specific word types falling into log frequency range Y.0 to Y.9 ÷ #word types found in frequency data-base X; with options for X being defined above and options for Y ranging from 1 to 9	1 (bands 1–6), 2 (bands 1–9), 3 (bands 1–6), 4 (bands 1–5)	✗

Table B.10: *Familiarity and informativeness features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Average Word Familiarity Per Million Words in frequency data-base X (all word tokens of type Y)	$\sum_i$ frequency per million word of word tokens in frequency data-base X that have the same length as word token *i* and start with the same three letters ÷ #word tokens in the text found in frequency data-base X; **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	2a–c, 3a–c
Average Word Familiarity Per Million Words in frequency data-base X (all word types of type Y)	Equivalent to `Average Word Familiarity Per Million Words in frequency data-base X (all word tokens of type Y)`	✗	2a–c, 3a–c
Average Word Informativeness Per Million Words in frequency data-base X (all word tokens of type Y)	$\sum_i$ cumulative frequency per million word of word tokens in frequency data-base X that have the same length as word token *i* and start with the same three letters ÷ #word tokens in the text found in frequency data-base X; **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	2a–c, 3a–c
Average Word Informativeness Per Million Words in frequency data-base X (all word types of type Y)	Equivalent to `Average Word Informativeness Per Million Words in frequency data-base X (all word tokens)`	✗	2a–c, 3a–c

Table B.11: *Age of active use features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Mean Age of Active Use in KCT (for word tokens of type Y)	(∑ age of all writers in KCT using this token of type X ÷ #word tokens of type X in the text found in KCT); **frequency data-base options for X:** see above; **word type options for Y:** see above	4a	4a–c
Mean Age of Active Use in KCT (for word types of type Y)	Equivalent to `Mean Age of Active Use in KCT (for word tokens of type X)`	4a	4a–c
Mean Age of Active Use in KCT for lemma types	Equivalent to `Mean Age of Active Use in KCT (for word tokens of type X)` but using only lemma types	4a	✗
Minimal Age of Active Use in KCT (for word tokens of type Y)	(∑ age of youngest writer in KCT using this token of type X ÷ #word tokens of type X in the text found in KCT); **frequency data-base options for X:** see above; **word type options for Y:** see above	✗	4a–c
Minimal Age of Active Use in KCT (for word types of type Y)	Equivalent to `Minimal Age of Active Use in KCT (for word tokens of type X)`	4a	4a–c
Minimal Age of Active Use in KCT for lemma types	Equivalent to `Minimal Age of Active Use in KCT (for word tokens of type X)` but using only lemma types	4a	✗

Table B.11: *Age of active use features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Maximal Age of Active Use in KCT (for word types)	(∑ age of oldest writer in KCT using this type ÷ #word types found in KCT)	✓	✗
Maximal Age of Active Use in KCT (for lemma types)	(∑ age of oldest writer in KCT using this lemma type ÷ #lemma types found in KCT)	✓	✗

B.4 Semantic complexity measures

This section contains definitions of all Semantic complexity measures (excluding language use features) that are used in this thesis.

Table B.12: *Semantic complexity features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Hypernyms per word type	#hypernyms found for a word type in GermaNet ÷ #word types found in GermaNet	✓	✗
Hyperonyms per word type	#hyperonyms found for a word type in GermaNet ÷ #word types found in GermaNet	✓	✗
Synsets per word type	#synsets found for a word type in GermaNet ÷ #word types found in GermaNet	✓	✗
Lexical units per synset	#lexical units ÷ #synsets	✓	✗
Relations per synset	#semantic relations ÷ #synsets	✓	✗
Frames per verb type	#frames ÷ #verb types found in GermaNet	✓	✗

B.5 Morphological complexity

This section contains definitions of all morphological complexity measures that are used in this thesis. Features are grouped into tables based on their complexity sub-domains. This section contains the following tables:

Table B.13 contains all measures based on the MCI.

Table B.14 contains all inflection-based morphological complexity measures.

Table B.15 contains all derivation-based morphological complexity measures.

Table B.16 contains all compound-based morphological complexity measures.

Table B.13: *MCI features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
MCI-10 for Nouns (± repetition)	MCI for nouns with X partitions à 10 exponents sampled with or without repetition for X ∈ { 2, 5, 10, 15 }	✗	✓
MCI-10 for Verbs (± repetition)	MCI for verbs with X partitions à 10 exponents sampled with or without repetition for X ∈ { 2, 5, 10, 15 }	✗	✓
MCI-5 for Adjectives (± repetition)	MCI for adjectives with X partitions à 5 exponents sampled with or without repetition for X ∈ { 2, 5, 10, 15 }	✗	✓
MCI-5 for Nouns (± repetition)	MCI for nouns with X partitions à 5 exponents sampled with or without repetition for X ∈ { 2, 5, 10, 15 }	✗	✓
MCI-5 for Verbs (± repetition)	MCI for verbs with X partitions à 5 exponents sampled with or without repetition for X ∈ { 2, 5, 10, 15 }	✗	✓

Table B.14: *Inflection features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Accusative Case per X	#accusative case markings ÷ X; **with X being: #nouns [1] or #words [2], finite verbs [3]**	1	2
Dative Case per X	#dative case markings ÷ X; with X being defined above	1	2
Genitive Case per X	#gentive case markings ÷ X; with X being defined above	1	2
Nominative Case per X	#nominative case markings ÷ X; with X being defined above	1	2
Any Person per word token	#1st, 2nd and 3rd person markings ÷ #words	✗	✓
First Person per token of type X	#1st person markings ÷ X; with X being defined above	3	2
Second Person per token of type X	#2nd person markings ÷ X; with X being defined above	3	2
Third Person per token of type X	#3rd person markings ÷ X; with X being defined above	3	2
Feminine inflection per word token	#feminine inflection markings ÷ #words	✗	✓
Masculine inflection per word token	#masculine inflection markings ÷ #words	✗	✓
Neuter inflection per word token	#neuter inflection markings ÷ #words	✗	✓
Gender inflection per word token	#feminine, masculine, and neuter inflection markings ÷ #words	✗	✓
Finite verbs per verb	#finite verbs ÷ #verbs	✓	✗
Non-finite verbs per verb	#non-finite verbs ÷ #verbs	✓	✗
Past participle verbs per verb	#verbs in past participle ÷ #verbs	✓	✗
Imperatives per Verb	#imperative inflections ÷ #verbs	✗	✓

Table B.14: *Inflection features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Imperatives per finite Verb	#imperative inflection ÷ #finite verbs	✓	✓
Imperatives per word token	#imperative inflections ÷ #words	✗	✓
Indicatives per Verb	#indicative inflections ÷ #verbs	✗	✓
Indicatives per finite Verb	#indicative inflections ÷ #finite verbs	✓	✓
Indicatives per word token	#indicative inflections ÷ #words	✗	✓
Subjunctives per finite Verb	#subjunctive inflections ÷ #finite verbs	✓	✓
Subjunctives per verb token	#subjunctive inflections ÷ #verbs	✗	✓
Subjunctives per word token	#subjunctive inflections ÷ #words	✗	✓
Imperfect tense per verb token	#imperfect tense marking ÷ #verbs	✗	✓
Past tense per verb token	#past tense marking ÷ #verbs	✗	✓
Number per word token	#singular and plural markings ÷ #words	✗	✓
Singular per word token	#singular markings ÷ #words	✗	✓

Table B.15: *Derivation features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Nominalizations of type X per token	#noun lemmas ending with nominalization X ÷ #words; **for X being: -ist [1], -eit [2], -ling [3], -keit [4], -at [5], -werk [6], -schaft [7], -enz [8], -tum [9], -ast [10], -eur [11], -ität [12], -ur [13], -heit [14], -keit [15], -nis [16], -wesen [17], -ator [18], -ismus [19], -atur [20], -ent [21], -ant [22], -arium [23], -ung [24],-ion [25]**	1–25	✗

Table B.15: *Derivation features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Derived nouns per noun	#derived nouns ÷ #nouns	✓	✗

Table B.16: *Compound features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Compound nouns per noun	#compound nouns ÷ #nouns	✓	✗
Compound depth per noun	∑#compounds in compound nouns ÷ #compound nouns	✓	✗

B.6 Discourse complexity

This section contains definitions of all discourse complexity measures that are used in this thesis. Features are grouped into tables based on their complexity sub-domains. This section contains the following tables:

Table B.17 contains all measures based on the use of connectives.

Table B.18 contains all measures based on the use of explicit co-reference.

Table B.19 contains all measures based on the use of implicit cohesive devices.

Table B.17: *Connective features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Additive connectives per Y (based on list X)	#additive connectives from list X ÷ Y; **with X being: list by Breindl [1] or list by Eisenberg [2] and Y being #tokens [a], #connectives on list X [b], #sentences [c]**	1c, 2c	1a
Adversative and concessive connectives per Y (based on list X)	#adversative and concessive connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a
Adversative connectives per Y (based on list X)	#adversative connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a
All connectives per Y (based on list X)	#connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a
Causal connectives per Y (based on list X)	#causal connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a

Table B.17: *Connective features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Concessive connectives per Y (based on list X)	#concessive connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a
Temporal connectives per Y (from list X)	#temporal connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a
Other connectives per Y (based on list X)	#other connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1a
Multi- to single-word connectives (based on list X)	#skip-/n-gram and unigram connectives from list X ÷ Y; with options for X being defined above	✗	1
Multi-word connectives per Y (based on list X)	#skip-/n-gram connectives from list X ÷ Y	1c,2c	1b
Single-word connectives per Y (based on list X)	#unigram connectives from list X ÷ Y; with options for X and Y being defined above	1c, 2c	1b
wenn-V1 conditionals per sentence	#conditionals formed with condition realized in a *wenn* (engl. "if")-clause and consequence in a verb-first clause ÷ #sentences	✓	✗
V1-*dann* conditionals per sentence	#conditionals formed with condition realized in a verb-first clause and consequence in a *dann* (engl. "then")-clause ÷ #sentences	✓	✗
wenn-dann conditionals per sentence	#conditionals formed with condition realized in a *wenn* (engl. "if")-clause and consequence in a *dann* (engl. "then")-clause ÷ #sentences	✓	✗

Table B.17: *Connective features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
V1-V1 conditionals per sentence	#conditionals formed with condition realized in a verb-first clause and consequence in a verb-first clause ÷ #sentences	✓	✗
Coverage of condition types	#condition clause types occurring at least once (V1-*dann*, V1-V1, *wenn*-V1, *wenn-dann*) ÷ #condition clause types measures ($N = 4$)	✓	✗

Table B.18: *Co-reference features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Pronouns of type X per token	#pronouns of type X ÷ #tokens; **with pronoun types X: all pronouns [1], personal pronouns (1st person) [2a], personal pronouns (2nd person) [2b], personal pronouns (3rd person) [2c], personal pronouns (any person) [2d], possessive pronouns (1st person) [3a], possessive pronouns (2nd person) [3b], possessive pronouns (3rd person) [3c], possessive pronouns (any person) [3d], person or possessive pronouns (1st person) [4a], person or possessive pronouns (2nd person) [4b], person or possessive pronouns (3rd person) [4c]**	1, 2a–d, 3a–d, 4a–c	✗

Table B.18: *Co-reference features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Pronouns of type X per noun	#pronouns of type X ÷ #tokens; with pronoun types for X being defined above	1, 2a–d, 3a–d	✗
Pronouns of type X per token in sentence per sentence	($\sum_i$#pronouns of type X in sentence i ÷ #tokens in sentence i) ÷ #sentences; with pronoun types for X being defined above	1, 2a–d, 3a–d	✗
Articles of type X per article	#articels of type X ÷ #tokens; with article types for X: definite [1] or indefinite [b]	1a–b	✗
Articles of type X per token in sentence per sentence	($\sum_i$#articles of type X in sentence i ÷ #tokens in sentence i) ÷ #sentences; with article types for X being defined above	1a–b	✗
Proper nouns per token	#proper nouns ÷ #tokens	✓	✗
Proper nouns per noun	#proper nouns ÷ #nouns	✓	✗
Proper nouns per token in sentence per sentence	($\sum_i$#proper nouns in sentence i ÷ #tokens in sentence i) ÷ #sentences	✓	✗

Table B.19: *Implicit cohesion features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as shorthand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Global Argument Overlap (lemma-based)	#argument lemmas overlapping between any two sentences in the text ÷ #sentences	✗	✓
Local Argument Overlap (lemma-based)	same as `Global Argument Overlap (lemma-based)` but calculated only between adjacent sentences	✗	✓

Table B.19: *Implicit cohesion features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Global Lemma Overlap	#lemmas overlapping between any two sentences in the text ÷ #sentences	✗	✓
Local Lemma Overlap	same as `Global Lemma Overlap` but calculated only between adjacent sentences	✗	✓
Global Lexical Overlap (lemma-based)	#lexical lemmas overlapping between any two sentences in the text ÷ #sentences	✗	✓
Local Lexical Overlap (lemma-based)	same as `Global Lexical Overlap (lemma-based)` but calculated only between adjacent sentences	✗	✓
Global Noun Overlap (lemma-based or type-based)	#nouns overlapping between any two sentences in the text ÷ #sentences	✗	✓
Local Noun Overlap (lemma-based or type-based)	same as `Global Noun Overlap (lemma-based or type-based)` but calculated only between adjacent sentences	✗	✓
Mean Global Argument Overlap (lemma-based)	#argument lemmas overlapping between any two sentences in the text ÷ #sentence pairs	✓	✓
Mean Local Argument Overlap (lemma-based)	same as `Mean Global Argument Overlap (lemma-based)` but calculated only between adjacent sentences	✓	✓
Mean Global Lemma Overlap	#lemmas overlapping between any two sentences in the text ÷ #sentence pairs	✓	✗

Table B.19: *Implicit cohesion features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Mean Local Lemma Overlap	same as `Mean Global Lemma Overlap` but calculated only between adjacent sentences	✓	✗
Mean Global Lexical Overlap (lemma-based)	#lexical lemmas overlapping between any two sentences in the text ÷ #sentence pairs	✗	✓
Mean Local Lexical Overlap (lemma-based)	same as `Mean Global Lexical Overlap (lemma-based)` but calculated only between adjacent sentences	✗	✓
Mean Global Noun Overlap (lemma-based or type-based)	#nouns overlapping between any two sentences in the text ÷ #sentence pairs; using noun lemmas [1] or noun tokens [2]	2	1
Mean Local Noun Overlap (lemma-based or type-based)	same as `Mean Global Noun Overlap (lemma-based or type-based)` but calculated only between adjacent sentences	2	1
Global Stem Overlap	#noun stems overlapping with stems of other lexical words in any other sentence ÷ #sentences	✗	✓
Local Stem Overlap	same as `Global Stem Overlap` but calculated only between adjacent sentences	✓	✗
Global Content Word Overlap (lemma-based)	#content word lemmas overlapping between any two sentences ÷ #sentences	✗	✓

Table B.19: *Implicit cohesion features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Local Content Word Overlap (lemma-based)	same as `Global Content Word Overlap(lemma-based)` but calculated only between adjacent sentences	✓	✗
Mean Global Stem Overlap	#noun stems overlapping with stems of other lexical words in any other sentence ÷ #sentence pairs	✓	✗
Mean Local Stem Overlap	same as `Mean Global Stem Overlap` but calculated only between adjacent sentences	✓	✗
Mean Global Content Word Overlap (lemma-based)	#content word lemmas overlapping between any two sentences ÷ #sentence pairs	✓	✗
Mean Local Content Word Overlap (lemma-based)	same as `Mean Global Content Word Overlap(lemma-based)` but calculated only between adjacent sentences	✓	✗
SD of Global/Local Argument Overlap (lemma-based)	standard deviation corresponding to `Mean Global/Local Argument Overlap (lemma-based)` feature defined above	✗	✓
SD of Global/Local Lexical Overlap (lemma-based)	standard deviation corresponding to `Mean Global/Local Lexical Overlap (lemma-based)` feature defined above	✗	✓
SD of Global/Local Noun Overlap (lemma-based)	standard deviation corresponding to `Mean Global/Local Noun Overlap (lemma-based)` feature defined above	✗	✓

Table B.19: *Implicit cohesion features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Transition probability from grammatical role A to grammatical role B	#transitions of noun heads (=entities) with grammatical role A to grammatical role B in a subsequent sentence ÷ (#sentences - 1) * #entities; **with options for grammatical roles of A being: subject [1], object [2], other complement [3], nothing [4] and options for grammatical roles of B being: subject [a], object [b], other complement [c], nothing [d]**	1a–d, 2a–d, 3a–d, 4a–d	✗
Propositional idea density	see Brown *et al.* (2008)	✓	✗

B.7 Processing complexity

This section contains definitions of all human processing measures that are used in this thesis.

Table B.20: *Human processing features used in this thesis (excluding raw counts). Check marks indicate if features were present in the legacy system and the CTAP system. # is used as short-hand to indicate counts of linguistic constructs*

Feature name	Definition	Legacy	CTAP
Sum longest dependency per sentence	(∑ max(#words in dependency per sentence)) ÷ #sentences	✓	✗
Sum longest dependency per t-unit	(∑ max(#words in dependency per sentence)) ÷ #sentences	✓	✗
Sum longest dependency per clause	(∑ max(#words in dependency per sentence)) ÷ #sentences	✓	✗
Sum longest dependency per finite clause	(∑ max(#words in dependency per sentence)) ÷ #sentences	✓	✗
Longest dependency	max(# words in a dependency)	✓	✗

Table B.20: *Human processing features used in this thesis (continued).*

Feature name	Definition	Legacy	CTAP
Maximal total integration cost per finite verb using configuration X	(∑ maximal total integration costs at the finite verb calculated using the configuration X) ÷ number of finite verbs; configuration options for X: *original weights (O), increased verb weight (V), decreased coordination weight (C), decreased modifier weight (M)* and weight adjustment combinations: *CV, CM, VM, CMV* Ex.: Maximal total integration cost per finite verb using CV weights (= decreased coordination weights and increased verb weights)	✓	✓
Total integration cost per finite verb using configuration X	(∑ total integration costs at the finite verb calculated using configuration X) ÷ number of finite verbs; configuration options for X: see above Ex.: Total integration cost per finite verb using CV weights (= decreased coordination weights and increased verb weights)	✓	✓
Adjacent high integration costs per finite verb using configuration X	(∑ adjacent integration costs > 2 after a finite verb calculated using configuration X) ÷ number of finite verbs; configuration options for X: see above Ex.: Adjacent high integration cost per finite verb using CV weights (= decreased coordination weights and increased verb weights)	✓	✓