Automatic Loanword Identification using Tree Reconciliation

Marisa Köllner, geb. Delz

Automatic Loanword Identification using Tree Reconciliation

Dissertation
zur Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von

Marisa Köllner, geb. Delz

aus

Künzelsau

2021

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Gerhard Jäger

Mitberichterstatter: Prof. Dr. Friedrich Hamm

Tag der mündlichen Prüfung: 12.03.2021

To Sven.

Thank you for being my rock in breaking waves.

Abstract

The use of computational methods in historical linguistics increased during the last years. Phylogenetic methods, which explore the evolutionary history and relationships among organisms, found their way into historical linguistics. The availability of machine-readable data accelerated their adaptation and development. While some methods addressing the evolution of languages are integrated into linguistics, scarcely any attention has been paid to methods analyzing horizontal transmission. Inspired by the parallel between horizontal gene transfer and borrowing, this thesis aims at adapting horizontal transfer methods into computational historical linguistics to identify borrowing scenarios along with the transferred loanwords.

Computational methods modeling horizontal transfer are based on the framework of tree reconciliation. The methods attempt to detect horizontal transfer by fitting the evolutionary history of words to the evolution of their corresponding languages, both represented in phylogenetic trees. The discordance between the two evolutionary scenarios indicates the influence of loanwords due to language contact. The tree reconciliation framework is introduced in a linguistic setting along with an appropriate algorithm, which is applied to linguistic trees to detect loanwords. While the reconstruction of language trees is scientifically substantiated, little research has so far be done on the reconstruction of concept trees, representing the words' histories. One major innovation of this thesis is the introduction of various methods to reconstruct reliable concept trees and determine their stability in order to achieve reasonable results in terms of loanword detection. The results of the tree reconciliation are evaluated against a newly developed gold standard and compared to three methods established for the task of language contact detection in computational historical linguistics.

The main aim of this thesis is to clarify the purpose of tree reconciliation methods in linguistics. The following analyses should give insights to which degree the direct transfer of phylogenetic methods into the field of linguistics is fruitful and can be used to discover borrowings along with the transferred loanwords. The identification of loanwords is a first step into the direction of a deeper understanding of contact scenarios and possible types of loanwords present in linguistic data. The adaptation of phylogenetic methods is not only worthwhile to shed light on detailed horizontal transmissions, but serves as basis for further, more detailed analyses in the field of contact linguistics.

Zusammenfassung

Die Verwendung von computerbasierten Methoden in der Historischen Linguistik stieg in den letzten Jahren stetig an. Phylogenetische Methoden, welche zur Bestimmung der Evolutionsgeschichte und Verwandtschaftsgraden zwischen Organismen entwickelt wurden, erhielten Einzug in die Historische Linguistik. Die Verfügbarkeit von maschinenlesbaren Daten förderten deren Anpassung und Weiterentwicklung. Während einige Algorithmen zur Rekonstruktion der sprachlichen Evolutionsgeschichte übernommen wurden, wurde den Methoden für horizontalen Transfer kaum Beachtung geschenkt. Angelehnt an die Parallele zwischen horizontalem Gentransfer und Entlehnung, werden in dieser Arbeit phylogenetische Methoden zur Erkennung von horizontalem Gentransfer für die Identifikation von Lehnwörtern verwendet.

Die Algorithmen für horizontalen Gentransfer basieren auf dem Vergleich zweier phylogenetischer Bäume. In der Linguistik bildet der Sprachbaum die Sprachgeschichte ab, während ein Konzeptbaum die Evolutionsgeschichte einzelner Wörter repräsentiert. Die Rekonstruktion eines Sprachbaumes ist wissenschaftlich fundiert, wohingegen die Rekonstruktion von Konzeptbäumen bisher wenig erforscht wurde. Eine erhebliche Innovation dieser Arbeit ist die Einführung verschiedener Methoden zur Rekonstruktion von stabilen Konzeptbäumen. Da die Algorithmen zur Erkennung von horizontalem Transfer auf einem Baumvergleich basieren, deuten die Unterschiede zwischen einem Sprachbaum und einem Konzeptbaum auf Lehnwörter innerhalb der Daten hin. Daher wird sowohl die Methodik, als auch ein geeigneter Algorithmus in einem linguistischen Kontext eingeführt. Die Ergebnisse der Lehnworterkennung werden mithilfe eines neu entwickelten Goldstandards evaluiert und mit drei weiteren Algorithmen aus der Historischen Computerlinguistik verglichen. Ziel der Arbeit ist zu erläutern, inwieweit Algorithmen basierend auf dem Vergleich zweier Bäume für die automatische Lehnworterkennung verwendet und in welchem Umfang Lehnwörter erfolgreich innerhalb der Daten bestimmt werden können. Die Identifikation von Lehnwörtern trägt zu einem tieferen Verständnis von Sprachkontakt und den unterschiedlichen Arten von Lehnwörtern bei. Daher ist die Adaption von phylogenetischen Methoden nicht nur lohnenswert für die Bestimmungen von Entlehnungen, sondern dient auch als Basis für weitere, detailliertere Analysen auf den Gebieten der automatischen Lehnworterkennung und Kontaktlinguistik.

Acknowledgements

I would like to express my deepest gratitude to everyone who supported me throughout this journey. Without all of you, it would have been impossible to complete this piece of work. Thank you.

I am truly and indebted grateful for my first supervisor Gerhard Jäger for his valuable guidance and support throughout this project. I thank him for providing me with an inspiring research field, while leaving me the freedom to explore this fascinating research topic. With his guidance and support, I managed to stay on the right path focusing on the important pieces of this research. Finally, I would like to thank him for his patience when parts of my work took longer than expected. In addition, I would like to thank my second supervisor Fritz Hamm for his guidance and support. He not only woke my interest in mathematics and inspired me to stay on track with the more formal parts of linguistics, but also provided me with insights into the mathematical foundations of the algorithms described in this thesis.

In particular, I would like to thank Johannes Wahle for being a great colleague and friend over the last years. We had an amazing time in Tübingen sharing an office, working and traveling together, and having insightful discussions. It was a pleasure working with you and I deeply thank you for supporting me in all occasions throughout this journey. I also thank Johannes Dellert and Taraka Rama for many instructive discussions and the development of new ideas on some of the research leading up to this thesis. I have learned a lot during our collaborations and this knowledge paved the way for a better realization of this project. I am grateful for all the discussions and feedback from Igor Yanovich and Christian Bentz. They always had time to listen to my ideas and gave me valuable feedback, also providing me with moral support when things did not work out as expected. In addition, I would like to thank Søren Wichman, Armin Buch, Bella Boga, and all other EVOLAEMP members for helpful discussions and feedback on my work during our time in Tübingen.

I am very lucky to have some awesome friends supporting me during this intensive time. A big thank you goes to Daniela Bentz for supporting me during the last years. Thank you for your deep friendship, lots of coffee, encouraging words, for always having an open door, and finally for the feedback on this thesis. This means the world to me. Another big thank you goes to Claudia and Enrico Reich. Without your support during the COVID-19 lockdown, it would not have been possible to finish this dissertation. Thanks Christl for having an open door and lots of coffee. Thank

you Mike for your deep and true friendship. Thanks Daniel for being a great friend and the best godfather for Ben. A big thank you goes to Sanni, Hendrikje, Rosa, Uli, and Susanne for your love, support, guidance, and for being my spiritual running buddies. Another big thank you goes to all my friends for their moral support and encouraging words.

All this would not have been possible without my family. My parents always have my back and keep supporting me no matter what. Thank you for being on my side and for encouraging me to find my own way. Thanks Mathea for being not only my sister, but also my best friend. I also would like to thank my parents in law, who provide me with a second home. Thank you for your endless love and support. A big thank you goes to my grandparents and grandparents in law. I can call myself lucky to have the support from all of you during this intense time. All of you have been great role models in order to find my own way and achieve my goals.

The greatest thanks of all goes to my husband Sven and my son Ben. I am deeply grateful to have you two in my life. You both are my rocks in breaking waves, my towers of strength. Thanks for always having my back, loving me during good and bad times, being patient during long working days, and for your endless love and support. The completion of this work would not have been possible without you. The two of you made this intense time to the best journey of my life. Thank you!

Contents

List of Figures

List of Tables

1 Introduction

Language is one of the greatest and fastest changing systems of mankind. The fascination of the evolution of languages along with their changes, contact scenarios, and the influence of surrounding factors inspires researchers to great innovations in the field of historical linguistics. Around the same time as Darwin (1871) presented his idea on the evolution of species and languages, Schleicher (1861) presented the first tree illustrating the classification of the Indo-European languages, along with an attempt to reconstruct their latest common ancestor: Proto-Indo-European. Based on the work of Eldredge (2005), Schleicher (1863) drew some parallels between the evolution of languages and species, indicating that the methods from natural science could be applied to linguistic data. Adopting the parallels of Schleicher (1863), Atkinson and Gray (2005) and Croft (2000) presented systematic overviews of the similarities between linguistic and biological evolution. In correspondence with those parallels, computational methods found their way from phylogenetics into linguistic research. The increasing availability of computational methods in bioinformatics and digitally accessible datasets in linguistics led to an application of the methods, resulting in a new research field: computational historical linguistics. While the most studied topics are the classification of languages, along with the detection of sound changes and cognates, less attention was paid to the process of borrowing. The similarities between borrowing and the biological process of horizontal gene transfer laid the foundation for the work presented in this thesis.

The transfer of words, also known as borrowing, is a horizontal transmission of linguistic items due to language contact. During the process of borrowing, the linguistic items are integrated and adapted into the recipient language, resulting in loanwords which are barely distinguishable from inherited words. In comparison, genes can be transferred between organisms without sexual reproduction and are fully integrated into the DNA of the recipient. This is known as horizontal gene transfer in biology. A popular example is the spread of antibiotic resistance in bacteria, where the genes responsible for the resistance are transferred from one species of bacteria to another through different processes of horizontal gene transfer, e.g. direct cell-to-cell contact.

The parallel between the evolutionary mechanisms cannot be denied. In recent years, phylogenetic algorithms are developed to detect horizontal gene transfer. The

methods are based on the framework of tree reconciliation, proposed by Goodman et al. (1979), which aims at detecting horizontal gene transfers by fitting the evolutionary history of genes to the evolution of their corresponding species, both represented in a phylogenetic tree respectively.

The aim of this thesis is the adaptation of horizontal transfer methods to test whether a direct adaptation is fruitful in terms of application, performance, and the results obtained from the algorithms. The approach of tree reconciliation highly depends on the input data, namely the reconstructed linguistic trees. One major innovation of this work is the introduction of various methods to reconstruct concept trees. Concept trees are equivalent to gene trees, i.e. they represent the evolutionary history of words within the tree model. Thus, the thesis tries to reconstruct reliable concept trees along with associated methods to determine their stability. In addition, the tree reconciliation framework is introduced in a linguistic setting and an appropriate algorithm for horizontal gene transfer is adapted and applied to linguistic data. The algorithm is compared to three methods developed in computational historical linguistics to detect language contact and borrowing scenarios. Finally, the results are evaluated against a newly developed gold standard containing all relevant information for an appropriate evaluation of loanwords and their underlying borrowing process.

During the first two chapters, all relevant concepts in historical linguistics and phylogenetics are introduced. In the third chapter, various methods for automatic concept tree reconstruction are presented. In a second step, resampling techniques are used to establish the reliability of the concept trees along with the reconstruction of tree replicates to generate tree samples. Finally, the methods are evaluated to determine the ideal approach for concept tree reconstruction, which is further used for the task of automatic loanword detection. The fourth chapter introduces four algorithms to automatically detect loanwords and their corresponding borrowing process. The phylogenetic tree reconciliation method can determine the loanword, the recipient language, the source language, and the direction of transfer, which is a main advantage over the other algorithms developed for this task. In addition, the tree replicates generated for the verification of the concept trees can be used by the algorithm to measure the stability of the detected transfer events. In the last chapter, all algorithms are evaluated against gold-standard data to get insights in the accuracy of their results and performance. The thesis concludes with a discussion and summary of the presented work and a short outlook on possible improvements and further research topics.

2 Historical Linguistics and the Computational Turn

Historical linguistics is a well-studied field for the evolution of languages along with their changes and contact scenarios. Various changes occur over time and affect a language in different ways, which leads to a continual development. On the one hand, there are internal changes, affecting the whole language and its varieties. For example, the simplification of verb forms in English and German, as illustrated by Delz et al. (2012). If one variety loses mutual intelligibility with another variety, it becomes a separate, unique language (François, 2015). On the other hand, a language can change drastically due to migration and contact with other languages. Those changes are difficult to discover, and a reconstruction requires time and knowledge about the language and its history. However, the evolution of a language can only be fully reconstructed if all changes and modification processes are known or can be discovered.

One of the main achievements in historical linguistics is the development of the *comparative method*. The comparative method is the central point of studies in language change and reconstruction. In contrast, the known phenomenon of language contact was long neglected in the research area of historical linguistics. The studies of contact-induced change led to a research field of its own, introducing a whole new set of questions to be answered. The research of *contact linguistics* focuses on contact situations, effects of contact, and the changes that come along in the languages under question. One process that takes place in a contact scenario of two languages is *borrowing*, which is the transfer of lexical items from one language to the other.

During the last decades, the interest of using computational models and analyses in historical linguistics increased. The known parallels between biology and linguistics caused the adaptation of phylogenetic methods in a linguistic setting. This computational turn opens up a new research area to answer questions which could not be answered using traditional linguistic methods. This includes, among other things, automatic language classification in a wide range of languages and language families, automatic cognate classification, identification of contact situations, and the effects and results of processes due to language contact, such as borrowing.

The first part of the chapter focuses on the main methods in historical linguistics,

like language classification, language contact, and borrowing. In the second part, the computational turn and the new insights in the field of historical linguistics that come along with the new methods are introduced. The priority lies on the methods needed for the detection of borrowings and loanwords, which are the main focus in the following studies.

2.1 Historical Linguistics

Historical linguistics is one of the oldest research fields in linguistics. Language comparison has a long tradition, starting back in antiquity, where scholars were aware of the similarities between Greek and Latin. Modern historical linguistics grew out of the field of philology in the late 18th century, at the latest with the first systematic study of diachronic language change from Jacob Grimm. His study on the development of the Germanic languages drew the attention to the research in historical linguistics and lay the foundations for *comparative linguistics*. Comparative linguistics addresses questions in the evolution, reconstruction, and classification of languages. According to Weiss (2015), language classification and the underlying *comparative method* can be dated back to Schleicher (1863), who introduced the first pedigree for Indo-European languages.
Around the same time, Schuchardt (1868) studied the mixture of languages. In his book *Vokalismus*, he made his famous statement: "There is no totally unmixed language"[1] (Schuchardt, 1868, p. 5). This book was the groundwork for studies in language contact and initiated a great amount of research in contact linguistics. During this time, language contact and the corresponding contact situations were studied with great interest. It was obvious that languages under contact exchange words. It is therefore not surprising that Paul (1886), and later Seiler (1907-1913), started to work on the classification of the transferred words. However, the first distinction between a *Lehnwort* (loanword) and a *Lehnprägung* (loan coinage or calque) was formulated by Betz (1949). Compared to a loanword, where the lexical item and its meaning are borrowed, a calque describes semantic borrowing, where only the meaning is transferred. Haugen (1950) further classified this framework into three categories: loanwords, loanblends, and loanshifts. This classification serves until today as the basis for loanword classifications.

[1]Translated from the original: "Es gibt keine völlig ungemischte Sprache" (Schuchardt, 1868, p. 5).

2.1.1 Language Classification

Languages which belong to the same language family are genetically related, which means that these languages derive from a common ancestor, a single *proto-language*. The classification of which languages are more related to one another within a family is called *subgrouping* (Campbell, 2013). Shared innovations are a reliable criterion for subgrouping and can be discovered with the help of the *comparative method*. The comparative method is the most important tool in historical linguistics to discover as much as possible from a proto-language by comparing the related languages in terms of shared innovations for the reconstruction.

The comparative method is a systematic process, following an order of different steps to accomplish the reconstruction of the proto-language (Campbell, 2013; Weiss, 2015). The first and most important step is the identification of shared innovations. Lexical items of related languages are compared with one another to establish *cognates* and *sound correspondences*. Cognates are words with a common etymological origin, which are used to determine regular sound correspondences between related languages. Sound correspondences or correspondence sets are sounds present in related words of cognate sets. The sounds correspond from one related language to another since they descend from a common ancestral sound. This mirrors the iterative character of the comparative method, since cognates and sound correspondences can only be identified together, which requires repetition of the comparison steps (Jäger and List, 2016). The lexical items under comparison should be present in all languages under question. Swadesh (1955) introduced a list of *basic vocabulary*, which was evaluated and both narrowed and extended in recent studies in the course of the establishment of different databases (see e.g. Holman et al. (2008) and Dellert and Buch (2018)). Basic vocabulary is said to be universal and stable against borrowing in all languages under investigation. The universality of the lexical items implies that most of the languages share a word for this meaning and cognates can be identified. In the classical comparative method, loanwords should be identified while establishing sound correspondences. Usually, but not always, the systematic sound correspondences found between cognates are not present in loanwords. The usage of basic vocabulary lists reduces this problem since borrowings are much rarer than in other parts of the vocabulary (Campbell, 2013). According to the cognates and sound correspondences found in words among the related languages, proto-sounds are reconstructed.[2] One of the guiding principles in sound reconstruction is that the majority wins (Campbell, 2013). This means if there is no evidence to the contrary, the sound in the correspondence set present

[2]For a detailed description of the comparative method and examples of sound comparisons and reconstructions, please see Weiss (2015) or Chapter 5 in Campbell (2013).

in the greatest number of daughter languages is picked for the reconstruction. The establishing of cognates and sound correspondences and the reconstruction of the proto-language leads to a subgrouping of the languages. The subgrouping clusters the languages by means of shared cognates and correspondence sets. The clusters can be represented in a tree diagram, which mirrors the classification of the languages.

Languages can be seen as unitary organisms, meaning they are related if they descend from a common ancestor. Relatedness is then analogous to that of animals or plants. Languages are related via lineages, which are created through sexual and asexual reproduction (Noonan, 2010). In historical linguistics, the unitary organism model was adopted in the early 19th century by Schleicher (1863).

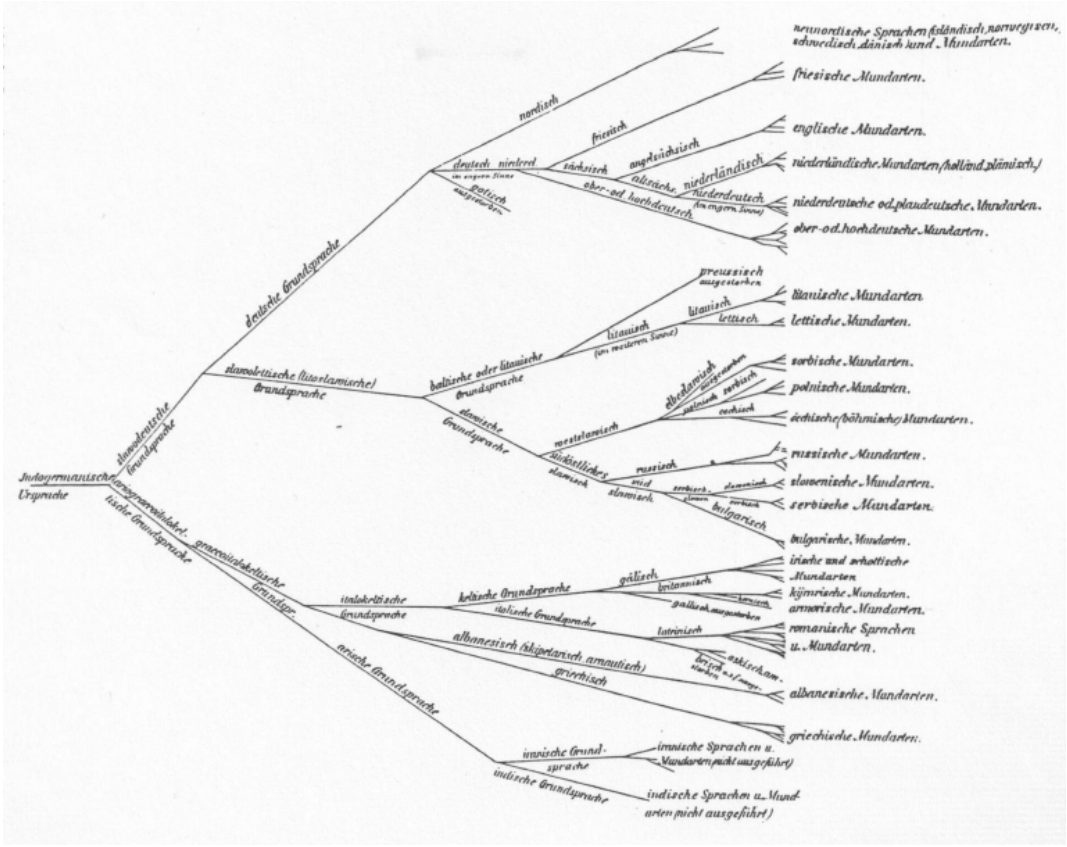


Fig. 2.1.: The family tree of the Indo-European languages from Schleicher (1863).

He was inspired by the work of Darwin (1871), since in one of his papers Darwin himself drew the parallel between biological and linguistic evolution. The family tree model is still preserved as the model of understanding and visualizing genetic

relations between organisms and languages (Noonan, 2010).[3] The comparative method is the underlying process to identify related languages, and groups them together into language families, which are represented in a language tree. The family tree model rests on a set of assumptions obtained from the unitary organism analogy. Following Noonan (2010), those assumptions are:

(i) Languages are unitary systems.
(ii) Languages are only genetically related if they descend from a single common ancestor.
(iii) New languages result from splitting off from an existing language.
(iv) A split is final and produces a new independent linguistic system.

In terms of understanding the history and relationships of languages, the issues coming along with these assumptions can be conceived, since none of the alternative models has been prevailing.[4] The tree model also assumes that language contact is irrelevant for establishing genetic relatedness, because the influence of contact cannot affect the genetic status of a language. This is in concordance with the comparative method, which assumes that language contact, and especially loanwords, have no impact on the comparison and reconstruction of languages. To reduce the noise in the analysis caused by loanwords, most studies aim at excluding them from the reconstruction and classification task. The comparative analysis, however, can also shed light on similarities of languages which are widely or not at all related to each other. Those similarities are mostly due to language contact and exchange of linguistic material, mostly lexical items. In the traditional comparative method, linguists can account for borrowing and loanwords by identifying them. Since in most cases loanwords would be excluded from the analysis, the effect of undetected borrowing events and loanwords on the classification of languages under the tree model remains unknown. As it can be seen in several computational studies (e.g. Holman et al. (2008), Jäger (2013b), and Jäger (2013a)), the effect is small and the subgrouping of the languages is still in accordance with trees reconstructed by expert linguists using the comparative method. However, since language contact cannot be traced in the tree model, it is still uncertain in what way borrowing influences the analysis. This is where contact linguistics comes into play, where other linguistic models were developed to describe language contact.

[3] Following the original analogy, the terms of mother/ancestor language, daughter and sister language are used in the linguistic family tree model (Noonan, 2010).
[4] For a summary of alternative models to the tree model, see Noonan (2010, p. 55).

2.1.2 Language Contact and Borrowing

Linguistic change can refer to different linguistic models. The theory of internal (sound) change is represented by the computational method and the tree model. The *wave theory*, an alternative to the tree model, is attributed to Schuchardt (1868) and Schmidt (1872), who were both students of August Schleicher. The wave model illustrates language change due to contact among languages, where each change is represented by a wave (Campbell, 2013). It starts from its region of origin and spreads over all languages sharing this innovation. Later changes may not cover the same area and can therefore lead to an overlap with already existing waves. This leads to the fact that there is no sharp boundary between the languages (Campbell, 2013). Figure 2.2 shows an illustration of the wave model.

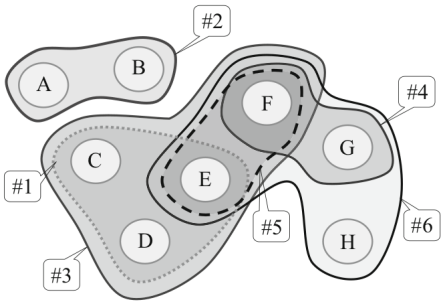


Fig. 2.2.: An illustration of the wave model, taken from François (2015). Languages are labeled A to H and innovations are numbered #1 to #6.

The first innovation (#1) is shared by languages C, D, and E, forming one wave. The second innovation (#2) is shared by languages A and B, and does not overlap with other waves and languages, e.g. this is the only change happening in those two languages. The other innovations are shared by several languages, each in some point overlapping with the other changes. In the case where two languages share an innovation, their similarity increases (e.g. E and F in innovation #5). If a change only affects one of the languages, their difference is increased and they diverge from one another (e.g. E and D in innovation #5). Many non-shared innovations can lead to a great distinction of the languages, and they can eventually become distinct (François, 2015). In the case of language variations, this can lead to a split into separate linguistic systems. Language subgroups are defined "as a group of languages whose ancestors participated together in the diffusion of one or several linguistic innovations, at a time when they were mutually intelligible" (François, 2015, p. 170). The wave structure can reflect a distribution of changes in intersected patterns, which is typical for dialect continua. For this reason, the wave model gained popularity in dialectology and was used to a lesser extent in contact

linguistics.

The complexity of generating a wave model of languages and display all possible contact situations starts with the difficulty of identifying contact scenarios between all languages under question. To get a real insight of the contact situations of languages, different components need to be considered. This includes, among others, social, geographical, and sometimes biological, cultural anthropological, paleoanthropological, and genetic components. Of course, not all components need to be known to reconstruct the contact scenario. However, they are of great interest to get a deeper understanding of the motivation of transferring linguistic items. Thomason (2010, p. 33) states:

> "It may reasonably be said that we have no full explanation for any linguistic change, or for the emergence and spread of any linguistic variant. The reason is that, although it is often easy to find a motivation for an innovation, the combinations of social and linguistic factors that favor the success of one innovation and the failure of another are so complex that we can never (in my opinion) hope to achieve deterministic predictions in this area."

Tendencies and probabilities can be found, but the question why one innovation can be adapted in one language but not in another under apparently similar circumstances cannot be answered.[5]

Considering only the linguistic component, language contact is defined as interaction of two or more languages with one another. In contact linguistics, the terminology is not clearly defined, and used differently by various authors. It is therefore important to offer some clarification. In general, a distinction is made between two categories of contact-induced changes: *borrowing* and *transfer* (which is synonymous to *interference* and *substratum influence*). The category of transfer can be divided further into subcategories, where additional categories are added by some authors to explain detailed changes.[6] The classification of different categories of contact-induced change makes it even more complex to categorize the various types of contact scenarios. Van Coetsem (1988) and Van Coetsem (2000) introduced a classification framework on contact-induced change. He defines two transfer types of cross-linguistic influence: *borrowing* and *imposition* (which is synonymous to *interference via shift* and *substratum influence*). The term *transfer* is used as a generic term referring to any kind of influence. A broad distinction between the two categories is that imposition is associated to second language acquisition and shift, whereas

[5] One of the major impulses for research in contact linguistics is the large-scale study of Thomason and Kaufman (1992). The study contains various contact situations with generalizations of language contact and its possible effects.

[6] A detailed description of the terminology in contact linguistics is given in Hickey (2012, p. 18).

borrowing is associated with language maintenance. If a French speaker uses French pronunciation while speaking English, it is imposition. During the process of borrowing, an English speaker would use French words while speaking English (Winford, 2010). Both processes require a source language (SL) and a recipient language (RL). The direction of the exchange is always from the source language to the recipient language. The distinction between the two processes is based on the psycholinguistic notion of dominance. The linguistically dominant language is the language in which the speaker is most proficient and fluent in. This does not need to be his native language (Winford, 2010). In the case of interference, the SL-dominant speaker transfers features into the RL. The RL can, for instance, be a second language (L2) which the speaker attempts to learn. In this case, the changes do neither affect the system of the RL nor of the SL. The changes can, however, spread in the language community of SL speakers learning the same L2. The RL spoken by its native speaker remains in most cases unaffected. In the process of borrowing, the RL-dominant speaker imports linguistic material from a non-dominant SL. Foreign material is adopted in the native language of the speaker, which modifies the language and its system (Winford, 2010).[7]

Borrowing is then defined as the acquisition of linguistic material from a SL into a RL via RL-dominant speakers, in other words, as the transfer of linguistic material from an external language into the language of the speaker. The extent of borrowing depends on the so-called *stability gradient of a language* (Van Coetsem, 1988). The stability varies between the different components of linguistic structure. Some domains of a language tend to be more stable and therefore more resistant to change than others are. More stable domains include syntax, semantics, phonology, and morphology (especially inflection paradigms). The lexicon and specific areas of structure, like derivational morphology and free function morphemes, are less stable and therefore receptive to change (Winford, 2010). Thomason and Kaufman (1992) combine the borrowability of linguistic material with the degree of contact. Intense contact can result in intense transfer, where no linguistic barrier to borrowing exists anymore. Lexical borrowing comes along with casual and slightly more intense contact, where content and function words are transferred between the languages, depending on cultural and functional reasons.[8] Lexical borrowing received considerable attention, since the probability of borrowing lexical items from one language to another is high. *Loanwords*, which are the result of the borrowing process, can be found in most languages of the world. The borrowing of lexical items does not only depend on the degree of contact between two languages, but also on both social and

[7] For a more detailed discussion of the distinction between the two processes, see Smits (1998) and Winford (2005).

[8] See Thomason and Kaufman (1992, p. 74) for a detailed borrowing hierarchy.

linguistic constraints.

The socially based motivation of borrowing is associated with need and prestige. Weinreich (1968, p. 56) noted that "the need to designate new things, persons, places, and concepts is, obviously, a universal cause of lexical innovation." Socially dominant and subordinate languages tend to borrow from each other. Depending on the social situation, both languages can exchange lexical items to fill the gaps for specific meanings. If one of the languages is socially dominant, e.g. during colonization, borrowing takes place from the subordinate language to the dominant one, but usually not the other way around. The need of borrowing increases in terms of technology, science, and higher learning in general, where a language wants to keep up with developments. In the Early Modern English period, English borrowed words from French, Latin, and Greek (Winford, 2010). Prestige, as a second social factor, is mirrored in the borrowings from French into English, where English extended the lexicon with words like *pork, beef, veal, etc.* to label forms of meat. German adopted the usage of the third person plural form as a polite pronoun form from French. French speakers used the different pronouns to express honor and respect to another person, which should create a gap between the higher and lower society. German adopted the concept of prestige and the different usage of the pronouns (Delz, 2013). However, to understand the motivations for lexical borrowing, further aspects of the contact between the speakers need to be taken into account. As Winford (2010) noted, "such factors [among others] include the patterns of social interaction between the groups, the degree of bilingualism, the demographics and power relationships, and attitudes toward the languages." The complete traceability of all social components of the borrowing process is complex due to the different scenarios involved, and is therefore the most challenging part. In most cases, there is no record of the social situation at the time of contact, and the socially based motivation for borrowing cannot be traced back at all.

Linguistic constraints on borrowing are based on borrowing hierarchies, which are split into two parts in terms of lexical borrowing: *morpheme hierarchy* and *part of speech hierarchy* (Haspelmath, 2008). The morphemic type expresses that lexical items are more likely to be borrowed than bound morphemes. Content words are therefore easier to borrow than function words, which is underlined by different linguistic studies (see e.g. Thomason and Kaufman (1992), Van Hout and Muysken (1994), Haspelmath (2008), and Tadmor et al. (2010)). In regard to parts of speech, Whitney (1881) already noted that nouns are borrowed easiest, and suggested the following hierarchy:

(2.1) nouns – other parts of speech – suffixes – inflections – sounds

The category of *other parts of speech* is not described in more detail. This hierarchy was extended by Haugen (1950) in his case study:

(2.2) nouns – verbs – adjectives – adverbs – prepositions – interjections

On the basis of English loanwords in Hindi, Singh (1981) proposed a slightly different hierarchy:

(2.3) nouns – adjectives – verbs – prepositions

All three borrowing hierarchies emphasize the widely acknowledged assumption that nouns are borrowed easiest. As Myers-Scotton et al. (2002) stated, nouns receive a meaning or a thematic role and do not assign it. Their insertion in another language does not disrupt the structure of the language. The order of the word forms on the borrowing scale depends highly on the data used for the case studies. The degree of typological distance between two languages can enhance or prevent the exchange of lexical items. For example, borrowing of verbs is facilitated if two languages have a similar verb system (Winford, 2010). This could be the reason why different case studies conclude different orders of the word categories. A systematic study on borrowings across languages can shed light on the arrangement of parts of speech in the borrowing hierarchy. The first systematic cross-linguistic research on borrowing was done by Tadmor et al. (2010) in the course of the *Loanword Typology project*. One of the results of the project was the *World Loanword Database* (WOLD), which was used to establish a borrowing hierarchy quantitatively.

(2.4) nouns – verbs – adjectives and adverbs

The WOLD database includes 41 languages. An overall word list size of 1,460 was specified, where each word was translated by specialists into the corresponding languages. During the translation process, each specialist identified loanwords in his/her corresponding language, which were marked as such in the database (Tadmor et al., 2010). According to this information, a quantitative study to establish a borrowing hierarchy could be carried out. The quantitatively established borrowing hierarchy in (2.4) reflects the general assumption that nouns are borrowed easiest, followed by verbs, adjectives, and adverbs.
One of the main questions in the identification of loanwords is, how inherited features can be distinguished from those which are borrowed. In the case of historical relatedness (regardless of type and source), it might not matter whether languages

share certain properties due to borrowing or inheritance, since both give evidence for historical connections between languages. One reason for this combination is the challenging task of identifying loanwords. Most of the time, one cannot be sure whether a word is a loanword, and at some point of the reconstruction of proto-languages, this question becomes unanswerable (Kessler, 2001). To address the question and face the challenging task of loanword identification, a better understanding of the different loanword types and their integration into the languages is needed.

2.1.3 Loanwords: Adaptation and Identification

The process of borrowing does not only include language contact and transfer of lexical items. The transferred words spread through the whole community of speakers and are integrated into the language system. This means each word is adapted according to pronunciation, inflection, and spelling to fit into the RL. Sooner or later, some of them even become indistinguishable from native words. Integrated loanwords behave like native words, and can therefore be subject to other processes of change. Before gaining insight in the process of integration and the identification of loanwords, a definition of loanwords is needed.
Haugen (1950) introduced three borrowing patterns to classify loanwords, which serve until today as the basis for loanword categorization.

Loanshifts *Loanshifts* include *loan translations* and *semantic loans*. Loan translations, known as *calques*, are often compounds which are recognized and translated by the speaker during the borrowing process. For example, the German word *Halbinsel* and the French word *presqu'île* are borrowed from Latin *peninsula*. Another example is the German word *Wolkenkratzer*, which is borrowed from English *skyscraper* (Haugen, 1950, p. 214). The compound is recognized by the speaker and both parts are translated into the RL.
A semantic loan is the result of a borrowing process where only a meaning is transferred and no structural or formal element. Haugen (1950, p. 214) uses the example of American Portuguese speakers using *humoroso* with the English meaning 'humorous', although it means 'capricious' in Portuguese.

Loanblends *Loanblends* are hybrids consisting partially of borrowed and native material. An example given by Haugen (1950, p. 219) is the word *bockabuch* (pocket book) in Pennsylvania German, where *bocka-* is borrowed from English *pocket* and

-buch is the German word for *book*. According to Haspelmath (2009), loanblends are not widely attested, since most of the lexical items which could be identified as such are not borrowings at all. Most of them are created based on loanwords (e.g. *desk lamp*, an English compound constructed using two Greek loanwords). Such a word cannot be counted as a loanword. This phenomenon makes it difficult to identify hybrids in a language without any further knowledge of the borrowing process or the contact situation.

Loanwords *Loanwords* are the most important types of lexical borrowings. The group of loanwords is the vaguest one, since it may include pratically the other two groups as well. Haugen (1950) limits it to words where the meaning, the form, and the phonemic shape, with more or less complete substitution, is integrated in the RL. Loanwords could further be specified according to their degree of substitution in the borrowing language. Most of the time, there is no record of the degree of integration or the adaptation process itself. Depending on the languages under question and the corresponding contact situation – including the point in time in the history, degree of contact, and the like – the understanding of the integration process is hardly traceable. However, in comparison to the other kind of borrowing patterns, loanwords are the most attested results of a borrowing process. Most of the research in lexical borrowing deals with the identification of loanwords and the reconstruction of the adaptation process.

Following Haugen (1950), Haspelmath (2008), and Winford (2010), a loanword is defined in the following way:

Definition 2.1 *A loanword is a word adapted from one language (SL) into another language (RL) at some point in the history of the RL, showing morphemic importation and substitution.*

An example are the following French loanwords in German, taken from Volland (1986).[9]

(2.5) a. Old French: raisin, rosin → borrowed into German: Rosine
b. Old French: pastee → borrowed into German: Pastete

[9]Further examples of loanwords in different languages can be found in Campbell (2013).

At this point in time, the words describing this specific kind of objects are absent in the German language and therefore borrowed from Old French. As it can be seen in the example, the words underwent an adaptation process in the German language to fit into the system. The integration involves the adjustment of sounds, phonological patterns, and syntactic properties, like number and gender, and fulfill the requirements stated in the definition of a loanword given in 2.1.
The modification of sounds takes place at the beginning of the borrowing process. In the traditional view, bilingual speakers who borrow words containing foreign sounds immediately change them. Due to *phonetic interference*, foreign sounds are modified to comply with native sounds and phonological constraints of the RL. This process is called *adaptation* or *phoneme substitution*. Campbell (2013, p. 59) defines the term adaptation in the following sense:

> "In adaptation, a foreign sound in borrowed words which does not exist in the receiving language will be replaced by the nearest phonetic equivalent to it in the borrowing language."

A good example is the integration of Germanic loanwords in Finnish, given in Campbell (2013). The voiced stops *b, d, g* are absent in the Finnish language. If a Germanic loanword contained one of those, they were replaced with the closest phonetic counterparts, namely the voiceless stops *p, t, k*.

(2.6) a. Germanic **bardaz* → Finnish *parta* (beard)
b. English *humbug* → Finnish *humpuuki*

In the process of *accommodation*, loanwords with non-native phonological patterns are modified to match the phonological combinations of the RL. This includes deletion, addition, or recombination of sounds until the words fit into the structure of the RL.[10] Additionally, the morphology and syntax of loanwords need to be adjusted to match the RL system. This means categories like gender, number, case, and the like need to be assigned to the loanwords. This is, however, not always straightforward, depending on the languages under contact. If the languages are similar in their morphology and syntax, the assignment is in most cases obvious. If not, there needs to be a criterion on which basis the loanwords can be integrated into the system. This can be done by means of comparisons to native items or already integrated

[10] The reconstruction of the adaptation and accommodation processes of loanwords is the center of interest in studies using Optimality Theory (OT). An overall explanation of using the OT framework to remodel the integration process of loanwords can be found in Delz (2013). Research studies using OT to model the substitution processes in specific recipient languages are, among others, the ones by Silverman (1992), Vendelin and Peperkamp (2004), Rose (2012), and Paradis and LaCharité (1997).

loanwords, by using similar criteria like affixes, gender, or stress. Some languages use a standard procedure to assign loanwords into their system. For example, German loanwords of the category verb are all assigned to the class of regular verbs. Loanwords of the category noun, where the gender cannot be determined, receive neutral gender in German. As already stated, the outcome of borrowing varies according to the degree of contact. Depending on the length and intensity of the contact, the adaptation process can differ. For example, loanwords can introduce new sounds and cause changes in the phoneme inventory of the RL (*direct phonological diffusion*) (Campbell, 2013). This influences the adaptation process of loanwords and can lead to various integration patterns. Words borrowed at the beginning of the contact situation may undergo a different adaptation process than words borrowed after the changes in the phoneme inventory. Only naming one of many examples which can be caused by intense and long contact between languages. This leads to the fact that there are typical patterns for the substitution of foreign sounds in a language, but they do not have to be uniform over time and degree of language contact. Another fact is that the adaptation process of loanwords is language-specific, depending on morphological and phonological criteria. All processes of integration are of different length and happen under different circumstances. This means no two processes are alike.

In theory, the adaptation process seems understandable and verifiable. However, rarely, if ever, a speaker can be caught during the actual process of borrowing and adaptation, i.e. there is no record of a whole process. How can we identify loanwords and detect the direction of borrowing? Linguists identify loanwords by finding similar shapes and meanings across languages where a contact situation exists and where another alternative explanation is missing or cannot be identified (Haspelmath, 2009). A common reason for similarities between languages is common ancestry, which needs to be excluded during the identification process. A clear determination of a loanword can only be ensured by verifying both the original pattern and a possible donor language. This itself is a challenge. In most cases, the original word is a reconstructed form, if it is present at all. The possible donor language cannot clearly be identified, i.e. there might be more than one possible candidate. There is not enough record of the language histories and the relatedness of the languages under question, which might lead to a wrong verification. There are family-internal borrowings which do not fulfill the above-mentioned criteria to clearly identify loanwords. Additionally, this process assumes that we already have possible candidates which could be loanwords in a language. Since loanwords can be fully integrated in a language, they cannot be identified without any knowledge about the language history, the possible contact situation, and the reconstruction of the language and its words. Only then assumptions can be made which words show

a different phonological pattern or are uncommon in a language. Campbell (2013) stated different criteria which can be assumed to clarify if a word is a loanword. The three strongest criteria are listed below.[11]

Phonological Clues *Phonological clues* can be found in words without a complete substitution. Sounds which are normally not expected or even absent in the RL are indicators for loanwords.
Furthermore, violation of the typical phonological patterns implies that words might be borrowed. Phonological patterns are syllable structure, morpheme structure, phonotactics, and the like. An example are the Mayan languages, having monosyllabic roots of the form CVC (consonant – vowel – consonant). All morphemes which violate this pattern turned out to be loanwords or compounds.
The knowledge of the phonological history of the language, including the reconstruction of the phoneme inventory and sound changes, can be helpful to detect loanwords, the direction of borrowing, and the donor language.

Morphological Complexity *Morphological complexity* can help to discover the direction of borrowing. Morphological complexity means a word is composed out of two or more morphemes, or has a morphologically complex etymology. The direction of borrowing is mostly from the morphologically complex language into the language with the monomorphemic word.
English borrowed the Spanish word *el lagarto* as *alligator*. In Spanish, the word is morphologically complex, since it consists out of an article and the noun itself. In English, after the borrowing process, the loanword has lost this complexity. An example of a morphologically complex etymology is the borrowing of the French word *vinaigre* into English as *vinegar*. The French word is a compound *vin-* 'wine' + *-aigre* 'sour'. The compound structure is absent in English, the individual words fuse into one noun.
This is a strong but complicated criterion, since the etymology of words is not always known, which could lead to false positives.

Cognates *Cognates* can shed light on the relatedness of words between languages. Words are subject to borrowing if cognates can be found in languages of another language family and not in most of the sister languages. The amount of attested cognates in the languages and its families indicates the direction of borrowing. The language family where the cognate is attested in most of its daughter or sister

[11] For detailed examples, please have a look at Campbell (2013, pp. 61–65).

languages serves as donor language, and the language(s) where the cognate is present in less daughter or sister languages serves as recipient language. The Finnish word *tytär* 'daughter' has no cognates in other Finno-Ugric languages. In Proto-Indo-European, *dhugəter 'daughter' has attested cognates in most Indo-European languages. Therefore, Finnish borrowed the word from an Indo-European language, namely Baltic (Campbell, 2013, p. 64).

This criterion leads to different challenges. First, there are languages where the cognate information is absent or not established, which results in incomplete data. The direction of borrowing could be identified wrongly depending on the missing cognates, but the loanwords could still be identified. If the cognate information is missing in the sisters of the RL, the status of the word as loan is questionable and needs to be verified using additional criteria. Second, older borrowings are inherited by the daughters of the language where the words belong to the same cognate set. The direction can therefore be unclear if there are enough languages that share the information. Additionally, it might be the case that a word would be seen as inherited instead of borrowed, since it shares the cognate set with all its sisters.

According to the definition of loanwords, it is clear that loanwords are opposed to native words. In his detailed description of how to define the term loanword, Haspelmath (2009, p. 38) examines the relation between the conceptions of loans and native words:

> "[...] Given our definition of [the term] loanword [...], we can never exclude that a word is a loanword, i.e. that it has been borrowed at some stage in the history of the language. Thus, the status of native words is always relative to what we know about the history of a language. English *dish* goes back to Old English and has cognates in other Germanic languages (e.g. German *Tisch* 'table'), so in this sense it could be regarded as a native word (contrasting with *disk*, which was borrowed from Latin *discus* in the 17th century). But we know more about the history of English than the attested forms in Old English: Proto-West Germanic **disk* has itself clearly been borrowed from Latin *discus*, so that English *dish* must count as a loanword after all. Even for words that have been reconstructed for a very ancient proto-language, such as English *mother* (from Proto-Indo-European *mātēr) or ten (from *dekḿ), we cannot be sure that they were not borrowed from another language at some earlier stage. Thus, we can identify loanwords, but we cannot identify "non-loanwords" in an absolute sense. A "non-loanword" is simply a word for which we have no knowledge that it was borrowed."

The term “non-loanword” is equivalent to native or inherited word. However, there is a tendency in historical linguistics to label only those words which have cognates in related languages or can be reconstructed to some proto-language as native or inherited. Loanwords can therefore only be identified if we can verify that a word has another phonological pattern than other words in the RL; and/or if the word belongs to a cognate set absent to most of its sister languages; and/or is present in other (unrelated) languages. This is in line with the statement of Haugen (1950), who noted that a double comparison is needed to identify borrowings. First, earlier and later states of the words in the given language need to be compared. Second, a comparison between the word under question and similar words in other languages is required.

A word comparison between different languages can be done in two manners. For the detection of loanwords in a specific language, words can be identified as loans either by the knowledge of linguists, or according to the phonological patterns. If the history of the language is known, most contact situations can be reconstructed. It is therefore straightforward to define a set of languages which could function as source languages. For a systematic cross-linguistic comparison of loanwords, a set of languages and words needs to be defined. The set of words should consist of lexical items present in most of the languages under question, where some of them are loanwords. The research in lexicostatistics provides word lists of universal items which are used for language classification. Even though basic vocabulary lists are said to be stable against borrowings, quite a few exceptions can be found. McMahon (2010, p. 131) summarizes the loanwords identified by Kessler (2001) across the 100 and 200 word Swadesh lists: 31 loans are identified for English, Albanian has a total of 41 loans (35 of these from Latin), French 27, and Turkish 22. Additionally, Embleton (1986) identified 24 English loanwords, 12 from North Germanic and 12 from French, and 15 Frisian loanwords from Dutch. This weakens the assumption that basic vocabulary is stable against borrowing, and draws the attention of identifying loanwords in basic meaning lists.[12] Using basic vocabulary lists simplifies the construction of word lists for a set of languages, since both databases and studies in comparative linguistics provide word lists across different languages (see e.g. Swadesh (1955), Kessler (2001), Holman et al. (2008), Haspelmath and Tadmor (2009b), Tadmor et al. (2010), and Dellert and Buch (2018)). The list of basic meanings for a set of languages allows for word comparisons to determine similar forms across related and unrelated languages.
The comparison between the earlier and later states of a word is a challenging

[12]See Haspelmath (2009, p. 48) for an explanation on the reasons of core borrowings.

task, since all words under question need to be reconstructed. The comparative method can shed light on the relatedness of words and their languages by identifying cognates and sound correspondences. Since cognates are opposed to loanwords, the detection of cognates and their corresponding cognate sets can provide an indication of a word being a loanword. Loanwords tend to be grouped together with similar words belonging to unrelated languages, and not together with their sister languages. In some cases even the direction of borrowing can be detected with the help of cognate sets, i.e. the language containing the loanword is grouped together in a cognate set with languages from another language family, which indicates the source of the word. For example, all Germanic words for the meaning 'mountain' are grouped together as cognates, except the English word *mountain*, which is grouped together with the Romance languages, indicating that the English word is a loanword. In this case, the direction of borrowing from the Romance language family to English can be identified, since English is the only Germanic language in the group of Romance languages. However, the specific source language within the sample of Romance languages cannot be determined by cognate classes. Sound correspondences and shared phonological patterns can be used for both detecting loanwords and their particular source language. The comparison of phonological patterns across languages can give insight about the similarity of words between languages. If languages share phonological patterns, this could either be an indicator of common ancestry or borrowing. The two possibilities can be distinguished by using the information of the languages' classification. If most of them share the same phonological pattern, the word is inherited. Otherwise, there is a high probability that it is a loanword. The direction of borrowing can be detected if the set of languages includes other languages with the same pattern which could function as donor languages. Using the example of the English word *mountain (*maʊntɪn*)*, according to phonological patterns the word is more similar to French *montagne (*mɔtaɲ*)*, Spanish *montaña (*montaɲa*)*, and Portuguese *montanha (*motɐɲɐ*)* than to Germanic languages like German *berg (*bɛɐ̯k*)*, Swedish *berg (*bɛrːɡ*)*, and Dutch *berg (*bɛrx*)* (Dellert, Jäger, et al., 2017).[13] The English word would be identified as loanword and French would function as donor language, since the phonological pattern of English and French is most similar. As it is known from the history of the English language, the native word for 'mountain' in Middle English was *berwe/bergh*, which was displaced by the French word *mountain* (Merriam-Webster Dictionary, 2020). Needless to say, in most cases the identification is not straightforward; however, the usage of the comparative method for the identification of loanwords via cognates and similar phonological patterns is a good starting point.

[13] Transcriptions are taken from the online database, see Dellert, Jäger, et al. (2017).

Summary: Loanword Identification as challenging task

The identification of loanwords, their adaptation process, and the detection of donor languages are challenging tasks. The adaptation during the borrowing process is language-specific, where no universal assumptions can be made. If possible candidates of loanwords are known, traditional linguistic theories, like Optimality Theory, can be used to reconstruct the adaptation process. The most promising method to identify loanwords and detect their source language is a systematic cross-linguistic comparison between a set of languages using historical methods. The comparative method functions as the underlying method for the detection task. The determination of cognates and sound correspondences can serve as criterion to distinguish an inherited word from a loanword. If the word is recognized as loanword, the similarity between phonological patterns can lead to the identification of the source language. The key point for all studies is a clear definition of loanwords, where it might be needed to ignore other (sub)categories to keep it a manageable survey.
For visualization purposes, the wave model needs to be rejected, since neither individual loanwords nor the direction of borrowing can be analyzed or displayed. A reconstructed language tree is a good basis for further development, since it represents the classification of the languages under question. In the last years, *networks* have been introduced in the research on language contact and borrowing, which combine language classifications with contact scenarios. However, the usage of networks causes further challenges, as it will be shown in the following chapter.
Additionally, it should be noted that only linguistic constraints on borrowings are taken into account. The borrowing hierarchy indicates that nouns are most likely to be borrowed. Working with basic word lists already implies this constraint, since most words are nouns, followed by verbs, and larger word lists contain also adjectives and adverbs. Including social constraints on borrowing in this kind of linguistic survey bears a different and probably even more difficult undertaking.
A systematic cross-linguistic study on borrowing and loanword detection is a challenging task, which can barely be done manually by linguists. It is time-consuming. The more languages in the set, the more time is needed to compare each word in each language with one another. Cognates, loanwords, and sound correspondences need to be determined, not to mention the detection of the donor language of each loanword. Fortunately, phylogenetic methods found their way into historical linguistics and unveiled a new field of study: computational historical linguistics. Computational methods are helpful tools to automatically detect cognates and sound correspondences, and to reconstruct language trees. And, as it will be shown in this research, they can shed light on the detection of loanwords and their source languages.

2.2 Computational Historical Linguistics

Computational approaches have been used in historical linguistics over half a century. During the last decade, the research in computational historical linguistics (CHL) experienced rapid growth due to the digitally available databases and the transfer of methods from phylogenetics.
Phylogenetics goes back to Darwin (1871), who introduced the tree model to study evolutionary relationships between species and organisms. The field of *cladistics* arose out of the methods for grouping species and organisms according to their common ancestry (Lecointre, 2006). These methods, which serve as basis for manual and computational phylogenetic analysis, were developed by the German entomologist Willi Hennig (Hennig, 1999).[14]
In his work, Darwin (1871) already mentioned the parallels between species and language evolution, which inspired Schleicher (1863) to his work on the Darwinian theory in linguistics. Based on this work, a systematic overview of the parallel types and mechanisms between biology and linguistics was outlined in the work of Croft (2000) and the study of Atkinson and Gray (2005). It is therefore not surprising that linguists transferred (computational) ideas and methods from phylogenetics. The research in historical linguistics benefits from the newly introduced approaches, since computational models allow for faster and larger data processing. In phylogenetics, methods of sequence comparison, taxon clustering, and tree reconstruction are already established and found their way into linguistic research. The following chapter introduces the computational elaborations of the comparative method and other computational approaches to study language contact, borrowing, and loanword detection.

2.2.1 Comparative Method: The Computational Turn

In cladistics, methods are elaborated to group species and organisms according to their ancestry. This is comparable to the aspects of the comparative method, i.e. the identification of sound correspondences and cognates to group genetically related languages. Further, the comparative method aims at detecting sound laws and reconstructing a proto-language and a tree diagram to represent the classifications of the languages. Jäger and List (2016) showed that a complete implementation of the traditional comparative method is currently not feasible. Some aspects correspond to phylogenetic approaches, for which computational methods are already established.

[14] The original book was written in 1966. In 1999, it was translated into English and published by the University of Illinois Press.

The systematic process used in the computational variant of the comparative method would include the following steps:[15]

1. Collection of basic vocabulary for a set of languages;
2. Sequence comparison to identify shared (phonological) patterns;
3. Cognate identification and grouping of cognate sets;
4. Reconstruction of a tree to classify languages.

A list of basic meanings can be extracted from different digitally available databases, which are introduced later in the chapter. Most of the databases focus on phonological presentations of the lexical items. Sounds are the basis for sequence comparison in the classical comparative method to establish sound laws and reconstruct proto-languages. In phonology, the *International Phonetic Alphabet* (IPA) is the standard representation of sounds from spoken languages (Yallop and Fletcher, 2007; O'Grady et al., 2011; International Phonetic Association, 2019). The advantage over orthographic forms is that all words are described with the same set of characters independent of their writing system. Additionally, some orthographically ambiguous words can be disambiguated using their phonological representation. Most of the databases therefore collect phonological representations of the basic vocabulary using IPA or a reduced form of it. This complies with the task of systematic cross-linguistic comparisons, since there is no barrier for the comparison of lexical items across different languages having different alphabets.
The detection of cognates and sound correspondences is a particular task of *sequence comparison*, which allows to determine shared patterns between the phonological representations of the lexical items. The identification of matching sounds due to word comparisons is the core analysis of the classical comparative method to establish sound correspondences in cognate pairs (Jäger and List, 2016). In phylogenetics, alignment analyses are a general task to model similarities and differences between two or more strings. Alignments are represented in matrices, where similar sounds are aligned in the same column and the empty cells are filled with gap symbols to indicate the differences between the strings. The computational alignment algorithms developed in biology and computational science are used in historical linguistics for a cross-comparison of phonological sequences. Generally, a distinction between two different alignment analyses is made: *pairwise phonetic alignment* and *multiple phonetic alignment*. Pairwise phonetic alignment is the comparison of two strings, resulting in an alignment matrix and a similarity or distance score. The following

[15] See Jäger and List (2016) and Jäger (2019) for a visualization of the steps. For a detailed technical description of all computational steps of the comparative method to receive language phylogenies, see Jäger and List (2016) and Dunn (2015b).

alignment example for the meaning 'mountain' shows automatic pairwise alignments of phonological strings.[16]

(2.7) Pairwise alignment of the words for the concept 'mountain':

a.
French: m o - t a 5 -
Spanish: m o n t a 5 a

b.
English: m a u n t - i n
French: m - o - t a - 5

c.
English: m a u n t i - n -
Spanish: m o - n t - a 5 a

d.
English: - m a - - u n t i n
German: b - E r k - - - - -

e.
German: b E r k
Dutch: b E r x

The alignments of English, Spanish, and French show a greater similarity than the alignments between English and German. This underlines the observation that the English word is cognate to the words in the Romance languages and therefore a loanword. In comparison, the German and Dutch word correspond with each other, which is indicated by the alignment without using any gap symbol. The additional computed similarity or distance score by the algorithm is a good indicator for the relatedness of languages and can be used for further processing. Multiple phonetic alignment, on the other hand, compares multiple strings with one another to construct an overall alignment matrix, which pictures the diversity within a cognate set. In example 2.8, the Swedish word *berg* is added to the alignment of Germanic languages showing the sound correspondences of the cognates. In the cognate class of the Romance languages, the diversity of the sounds within the cognate set is illustrated.

(2.8) Multiple alignments for the concept 'mountain':

a.
German: b E r k
Dutch: b E r x
Swedish: b E r g
...

[16] For representation purposes, the words are transcribed using the reduced phonological characters introduced in Brown et al. (2008) from the ASJP database (Wichman et al., 2018). 3 = high and mid central vowel, rounded and unrounded (IPA: ɨ, ɘ, ə, ɜ, ʉ, ɵ, ɛ̈). 5 = palatal nasal (IPA: ɲ).

b.
English: m a - u n t i - n -
French: m - o - - t - a 5 -
Spanish: m - o - n t - a 5 a
...

The multiple alignments can differ compared to the pairwise alignments depending on the algorithm and the underlying technique. It is therefore advisable to test different phylogentic algorithms on linguistic data to determine the most suitable one.[17]

The assessment of sound correspondences leads to the assignment of cognate classes, which are determined according to the alignments and established sound correspondences. Words with the same cognate class form a cognate set. The grouping can be represented in a binary presence-absence matrix, where 1 indicates the presence and 0 the absence of the cognate class.

		cognate sets		
language	phono. string	m	b	f
French	mota5	1	0	0
Spanish	monta5a	1	0	0
English	mauntin	1	0	0
German	bErk	0	1	0
Dutch	bErx	0	1	0
Swedish	bErg, fyEl	0	1	1
Norwegian	bErg,fyEl	0	1	1
Icelandic	fEtl	0	0	1

Tab. 2.1.: Binary (presence-absence) coding of cognate classes for the concept 'mountain'. The abbreviations stand for: Germanic languages b = 'berg', f = 'fjäll'; Romance languages: m = 'mountain'.

For illustration purposes, the Germanic languages containing a variation of the word *fjäll* for the concept 'mountain' are added to the example in table 2.1. All present Romance languages and English are assigned to the cognate class m, whereas German and Dutch are grouped together in b. This is in accordance with the alignment in example 2.7. The Germanic languages containing another word for the concept 'mountain' are grouped together in f. Since Swedish and Norwegian have both two words referring to the meaning, they are grouped in two cognate classes respectively. As it can be seen, the task of *automatic cognate detection* is a clustering or partitioning task, since the goal is a grouping of phonological sequences according to their similarity. Automatic methods used in historical linguistics promise good results,

[17] See Jäger and List (2016) for a detailed description of the alignment methods and an overview of computational alignment algorithms used and developed in historical linguistics.

which are confirmed in several studies (see e.g. Dunn et al. (2011), Bouckaert et al. (2012), List (2014), Rama et al. (2017)).[18] Some historical linguists provide cognate information for their data, like Ringe et al. (2002) or *The Austronesian Basic Vocabulary Database* (ABVD) by Greenhill et al. (2008). The cognate judgments in the databases can either be used for further analyses, or to evaluate the automatically detected cognates and their groupings.

In the comparative method, the languages are clustered according to the shared innovations. The more innovations two languages share, the closer they are related and clustered in the tree. The similarity/distance scores or the cognate classes can be used to reconstruct a language tree using phylogenetic methods. There is a wide range of algorithms developed in computational biology, and most of them are applicable out of the box in historical linguistics.[19] Both phylogenetics and historical linguistics strive for the same goal using tree reconstruction, namely the representation of observed languages at the leaves of the tree. The internal nodes of the tree should present inferred historical stages of the languages and their families. Noonan (2010) mentioned the assumptions made with the tree model, which are all reflected in the tree diagram reconstructed from automatic methods. The internal nodes should represent the common ancestor of two languages or groups of languages, indicating a language family. Further, the internal nodes display the split of two languages into unitary systems, which causes a new independent linguistic system. The automatically reconstructed tree in figure 2.3 displays the historical correct classification of the Germanic and Romance languages. Several linguistic studies demonstrate that compared to an expert classification tree, automatic tree reconstruction algorithms build trees with equally good language classifications (see e.g. Brown et al. (2008), Holman et al. (2008), Jäger (2013a), and Jäger (2013b)). The tree model assumes that language contact is irrelevant for the classification of languages according to shared innovations. The reconstructed tree in figure 2.3 displays the historically correct classification of the languages, although loanwords are present in the underlying data.

Depending on the language sample and the underlying data, loanwords introduce either a low or a high amount of noise to the analysis. The effect of noise varies according to the loanwords present in the basic meaning lists of the languages under question. In figure 2.3, the known loanwords introduce a low amount of noise, which leads to the correct classification of the languages. One of the reasons is that, on average, two languages share more cognates than they share loanwords. The cognate information between genetically related languages is more powerful than

[18] For a detailed description and overview of the current methods, see Jäger and List (2016).

[19] Delz (2013), Dunn (2015b), and Jäger and List (2016) give detailed explanations of the tree reconstruction methods in historical linguistics.

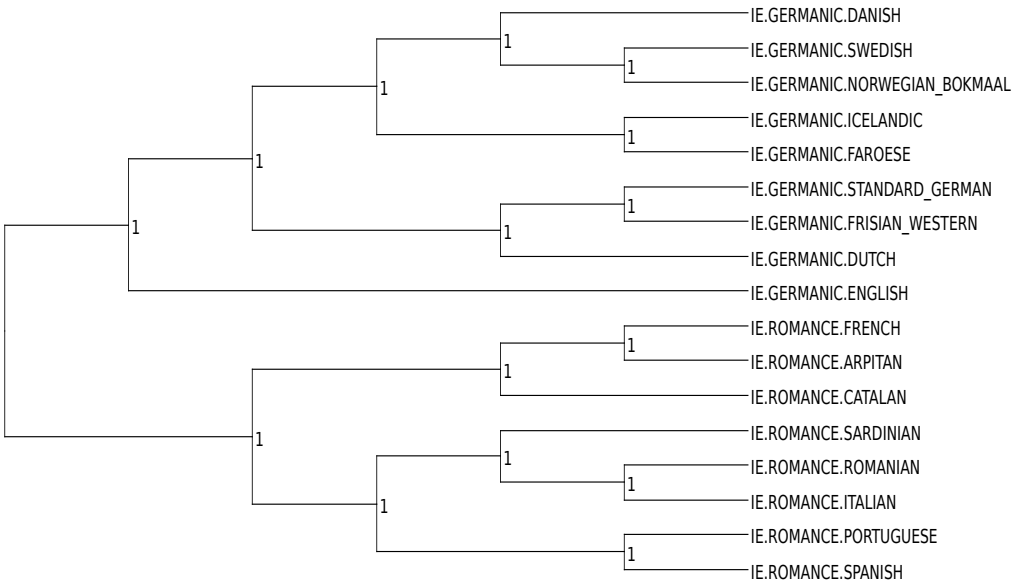


Fig. 2.3.: Automatically reconstructed tree of the Germanic and Romance languages from Delz (2014).

the loanword information. For example, Embleton (1986) identified 24 loanwords in English, 12 from French and 12 from North Germanic, in the 200 word Swadesh list. If those are the only loanwords in the list for English, the other 176 words should then be cognates. The 12 loans from North Germanic are harder to identify, since the languages belong to the same language family and might be grouped into cognate sets including other Germanic languages. If all French loanwords are grouped to the Romance languages, there are still 188 words which cluster English into the Germanic language family. It seems that in the sense of language tree reconstruction, loanwords carry less weight for a proper language classification. However, the amount of noise introduced to the analysis and its effect remains unclear if the correct number of loanwords cannot be established. If loanwords can be identified and excluded from the meaning list, the effect of noise in automatic tree reconstruction can be minimized. This could result in an increased consistency between the automatic reconstructed language trees and expert classifications. In addition, the detected loanwords are good indicators for language contact and borrowing, as it will be seen in the following studies.

2.2.2 First Attempts on Loanword Identification

Automatic methods for the detection of borrowings and loanwords are still in their infancy. Until now, there is no suitable method for this kind of identification task. Although the field of phylogenetics provides interesting methods and approaches, it seems that the transfer of these kind of approaches is more difficult than expected. One of the reasons is that, in classical historical linguistics, the task of loanword identification is already difficult enough, as it is previously shown. Language contact does not proceed in a regular manner, i.e. there are no universal constraints which could be applied for the identification, and both linguistic and social constraints play a major role during the process of borrowing. Not to mention the different kinds of loanwords which result from the borrowing process. If the identification of loanwords by classical methods is already a challenge, how could automatic approaches solve this task? First, there are limitations and concessions which need to be made while applying automatic methods from phylogeny. Second, different approaches need to be tested in order to shed light on the suitability of the algorithms. Third, the issue of data availability for training and testing data is still present. The problem in creating high-quality databases lies, on the one hand, on the time-consuming task of annotating and digitizing the data, and on the other hand, on the limited knowledge of the borrowing process between the languages of the world (List, 2019). In this chapter, attempts on loanword detection which do not fit in the later chapters are presented to illustrate the several ideas of automatic approaches.

One of the first attempts, to my knowledge, to identify loanwords via a formula was the study of Embleton (1986). The main aim of the study was, instead of excluding the loanwords from the word lists to compute language-relatedness, to introduce a borrowing parameter that can be added to the computation to correct for the effects of loanwords. The borrowing parameter is computed by identifying the number of loanwords in a RL from all its neighboring languages, resulting in one borrowing parameter for each language pair. The issues with this approach are that first, the parameter is language pair-specific. And second, McMahon and McMahon (2005, p. 91) showed that the number of loanwords needs to be known beforehand, i.e. the loanwords are identified manually. The method is therefore no help in the detection of unknown borrowings or loanwords across languages.

Minett and Wang (2003) propose a model using parsimony methods (Fitch, 1971) to detect conflicts of cognate sets on a given phylogeny. The cognate sets are distributed over seven Chinese dialects using multistate modeling of the lexical cognates, which excludes possible synonyms within a concept. In a later study, Wang and Minett (2005) introduced the idea of *skewing* to detect language contact, where the similarity between languages serves as underlying computation. However, the method

was not extended to detect loanwords if the result of the skewing method indicates borrowing between two languages.

In their work, Ringe et al. (2002) aimed at finding the "perfect phylogeny" for the Indo-European languages, which turned out to be more complicated than expected. The tree is reconstructed and refined using their own implementation of the maximum compatibility method. Wang and Minett (2005, p. 123) pointed out that although no single true tree could be established, the remaining incompatible characters could be indicators for non-genetic processes, like borrowing. Applications using the computational technique of *answer set programming* in combination with the implementation of Ringe et al. (2002) are introduced to model contact-induced innovations (Erdem et al., 2003; Brooks et al., 2005). The results look promising for small datasets, like the 24 Indo-European languages used by Ringe et al. (2002). However, for larger datasets, the computation is not feasible under the introduced models. In a follow-up study, Nakhleh et al. (2005a) extended the model of Ringe et al. (2002) to address the issue of incompatible characters due to contact. The aim of the study is to reconstruct a "perfect phylogenetic network" which consists of a single true tree containing a small number of contact edges. The contact edges are bi-directional, i.e. characters can be borrowed in both directions. Although the number of incompatible characters can be established, no effort was made to specify the loanwords or provide results about particular incompatible characters. One of the main issues is that the parsimony approach leads to more than one possible network, and the conflicts between optimization criteria for "perfect phylogenetic networks" cannot be solved (Nakhleh et al., 2005a).

The work on loanword detection of Van Der Ark et al. (2007) is based on sequence comparisons between phonological representations of lexical items. Similar to the procedure in the computational comparative method, phonological sequences are compared with a corresponding alignment algorithm to compute a distance score. The idea that loanwords have a small phonetic distance to words from which they were copied seems to be promising to detect loanwords. Both in the first study by Van Der Ark et al. (2007) and in the follow-up study by Mennecier et al. (2016), a near-zero word distance between unrelated languages is the indicator for loanwords. The main issue is the definition of a threshold by which words are judged to be loanwords. This is a crucial point using sequence-based methods for loanword detection tasks. Van Der Ark et al. (2007) and Mennecier et al. (2016) solved this issue by using annotated loanword data to determine the thresholds. It is unclear whether the established thresholds are achieving equally promising results for other language samples. Ideally, a method for loanword detection should not completely rely on expert loanword judgments to avoid circularity in the analysis. In addition, databases providing loanword annotations are only available to some extent.

2.2.3 Databases

The availability of digital databases increased during the last decades. On the one hand, some databases focus on one particular language family, including expert cognate judgments. Since the manual annotation of cognates is a time-consuming task, the focus on one language family is not surprising. Those databases are used for evaluation purposes of automatic cognate detection tasks. On the other hand, some databases focus on the coverage of a wide range of language families at the cost of expert cognate annotations. Those databases are mostly used to detect sound correspondences, where the computational methods benefit from the large amount of data.

While most of the time, the databases are distinguished according to the presence or absence of expert cognate judgments, in this thesis the main focus lies on the availability of loanword judgments. Except for the WOLD database, which was created for systematic cross-linguistic research on borrowing, existing databases are extended with loanword annotations. The choice of the database depends on the study and the research question of the loanword detection task. A database focusing on a particular language family can be used to detect family-internal loanwords, while with a database including languages from different families, family-internal and external loanwords could be identified. A presence-absence annotation of loanwords can only be used to verify the identified loanwords, whereas with the information of the donor language, the borrowing direction could additionally be evaluated.

In the following, the four main databases providing loanword annotations are introduced. This list does not aim for completeness, it rather presents the most prominent digital databases available with loanword annotations, and the one used throughout this thesis.

IELex The *Indo-European Lexical Cognacy Database* (IELex) (Dunn, 2015a) is based on an early study of Dyen et al. (1992), and contains basic vocabulary lists of Indo-European languages. The database contains a meaning list of 225 lexical items and their translations into the 24 languages provided by Ringe et al. (2002), plus additional 139 languages. The phonological representations are not presented in uniform IPA, i.e. many forms are transcribed in the Romanized IPA format introduced in Dyen et al. (1992). For each lexical item, the corresponding cognate class is determined according to the classifications of Dyen et al. (1992). During the last years, presence-absence annotations of loanwords were added to the database, where 1 means loanword and 0 either inherited or not annotated. The inconsistent

transcription of the lexical items requires post-processing of the data to use it with computational methods. Additionally, the presence-absence representations of the loanwords can neither distinguish between inherited and unannotated words, nor give an indication of the donor language.

ASJP The *Automated Similarity Judgment Program* (ASJP) (Wichman et al., 2018) is a cross-linguistic database covering over 5,000 of the world's languages in version 18. A reduced meaning list of 40 lexical items is used in order to achieve such a high coverage of languages. According to Holman et al. (2008), the 40 meanings on the list are enough to keep a stable classification of the languages on an automatically reconstructed tree. The lexical items in the database are transcribed using a 41 symbol encoding, out of which 7 symbols are for vowels and 34 for consonants (Holman et al., 2008).[20] In the last years, presence-absence loanword annotations have been added, where 1 indicates loanword and 0 inherited or missing annotation. Although the database provides a good coverage across the world's languages and probably enough data for automatic loanword detection, the distinction between inherited words or missing annotations is crucial for a statistical evaluation. Additionally, the donor languages of the loanwords cannot be extracted from the database, which are substantial for an evaluation of the direction of borrowing.

NorthEuraLex The *NorthEuraLex* (NELex) database (Dellert, Jäger, et al., 2017) is a large-scale lexicostatistical database generated within the *EVOLAEMP* project at the University of Tübingen.[21] A wide range of 1,060 concepts across 107 languages in Northern Eurasia is covered in the database. The unified IPA coding of the lexical items is generated automatically out of orthographies or standard transcriptions. The result is a constant transcription format across all languages, which is useful for all computational tasks, like language classification, cognate detection, and loanword detection. Additionally, the word lists can be converted automatically into the reduced ASJP encoding or other fine-grained transcriptions (Dellert et al., 2020). NELex is a large cross-family database for a well-researched linguistic area in the world, which is a great advantage in the task of loanword detection. In addition, the database is under constant development. Within the ongoing EtInEn project, expert cognate and loanword judgments are collected (Dellert, 2019b).[22] For loanwords, the donor language and a source word, if present in the source text, is provided. Further, the borrowing process and its history is reconstructed as far

[20] A detailed transcription of the alphabet can be found in the appendix of Holman et al. (2008).
[21] `https://www.evolaemp.uni-tuebingen.de/`
[22] `http://www.sfs.uni-tuebingen.de/~jdellert/etinen`

as the information is available.[23] Since NELex covers a wide range of concepts on languages from different language families and provides expert loanword judgments including the donor languages, this database is the one of choice for the following studies.

WOLD The World Loanword Database (Haspelmath and Tadmor, 2009b) is a result of the empirical study of borrowability in the collaborative Loanword Typology project (Haspelmath and Tadmor, 2009a). The database contains 41 individual languages, where for each language a specialist selects counterparts for items on a 1,460 fixed meaning list. WOLD provides information like orthographic and phonological form, as well as details about the loanword status, the donor language, and the source word with its meaning. The languages were selected "to represent the world's genealogical, geographical, typological, and sociolinguistic diversity" (Haspelmath and Tadmor, 2009a, p. 3). The diversity of the languages in the sample is well-suited for a systematic cross-linguistic study of loanwords and the borrowability of languages. However, for an automatic loanword detection approach, where the comparative method serves as basis, the diversity of the languages is obstructive. A good balance between related and unrelated languages is needed in order to get reliable results from the computational comparative method. Nevertheless, the loanword annotations and the corresponding donor languages can still be used for evaluation purposes. All loanwords where the meanings, the languages, and the donor languages are included in the NELex database are extracted and added to the evaluation set.

[23] The collection of the data is still under development. A preliminary version was kindly provided by my colleague Johannes Dellert.

3 Modeling Horizontal Transmission: Trees and Networks

The parallels between biological and linguistic evolution led to the adoption of phylogenetic methods to historical linguistics. While most of the studies in computational historical linguistics are concerned with language comparison, cognate detection, and language classification, less attention is paid to the parallels between *horizontal gene transfer* (HGT) and borrowing.

HGT (also called *lateral gene transfer*) is the transfer of genetic material between unrelated organisms without sexual reproduction (Morrison, 2011). A popular example from biology is the spread of antibiotic resistance in bacteria, where the genes responsible for the resistance are transferred from one species of bacteria to another through different HGT processes, e.g. direct cell-to-cell contact. Such a transfer of genes causes a change within an organism by adding new genetic material. This leads to a modification of the lineage, and the organisms' classification on the tree can be affected. However, similar to historical linguistics, the genes derived from sexual reproduction have a stronger impact on the tree than the transferred genes, i.e. the classification of the organisms in the tree is not affected by horizontal transfer. In biology, during sexual reproduction the offspring inherits genetic material from the parents. Due to recombination, the transferred genes are passed on to the next generation along with the inherited genes, and lead to a new mixture of the genetic material. With each reproduction process, the combination of the genes changes, and transferred genes are more and more integrated into the organism's structure of genetic material. The parallel can be drawn to borrowing, where words are transferred between languages without any form of inheritance and only through language contact. Words are the genes of the language and object to horizontal transfer between different languages. After the transfer, the words are passed on to the daughter languages along with the inherited words, and at some point in time, the words are completely adapted into the language.

In biology, horizontal transmission is modeled using networks to represent both the tree due to inheritance and the horizontal transfer of genes between organisms. Since a visualization of borrowing using the wave model was rejected in the last

chapter, networks offer a new and interesting opportunity to model borrowing, and explicitly loanwords. The field of phylogenetics provides a wide range of different algorithms to reconstruct different kinds of networks. Keeping the parallel of horizontal gene transfer and borrowing in mind, suitable methods and networks are applied for the purpose of modeling borrowing using the methodology of HGT.
Since different kinds of trees are a key principle for understanding the underlying methods of HGT and networks, those are introduced first in this chapter. Second, phylogenetic networks, their purpose, and their counterpart in linguistics are represented. Additionally, different kinds of networks to model language contact and borrowing used in linguistic studies are discussed to illustrate their usage and representations in linguistics. The last part of this chapter introduces the underlying methodology of horizontal gene transfer and the usability of this method in linguistics.

3.1 Phylogenetic Trees in Linguistics

In his famous book, Felsenstein states that "phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically" (Felsenstein, 2004, p. xix), which indicates the important role of tree structures in phylogenetics. In linguistics, the comparative method illustrates the importance of the tree concept to display language classifications. Until now, tree diagrams are the best concept in both fields to represent the relatedness of entities in a graphical way.
Phylogeneticists examine the tree from a mathematical point of view and have established several reconstruction algorithms. The tree itself is a versatile graph which is, on the one hand, flexible when dealing with different changes and manipulations, and on the other hand, informative according to the classification of the entities under question. Although a tree cannot represent all evolutionary processes, it can serve as fundamental idea or basis.
Trees can have different graph theoretic properties for the representation of the data, e.g. they can be either unrooted or rooted. The property of the tree illustrates a specific kind of classification of the entities under question. Depending on the underlying research question, the representation of the data and the corresponding illustration of the tree needs to be chosen. Furthermore, the labels on the tree can depict various kinds of data. Depending on the displayed data on the leaves, the interpretation of the trees differs, i.e. the histories of species and genes can both be represented using a tree.

The computational comparative method, introduced in section 2.2.1, outlines the computational steps for the reconstruction of phylogenetic trees. Figure 3.1 visualizes the main steps for data processing to obtain an automatically reconstructed tree.[1]

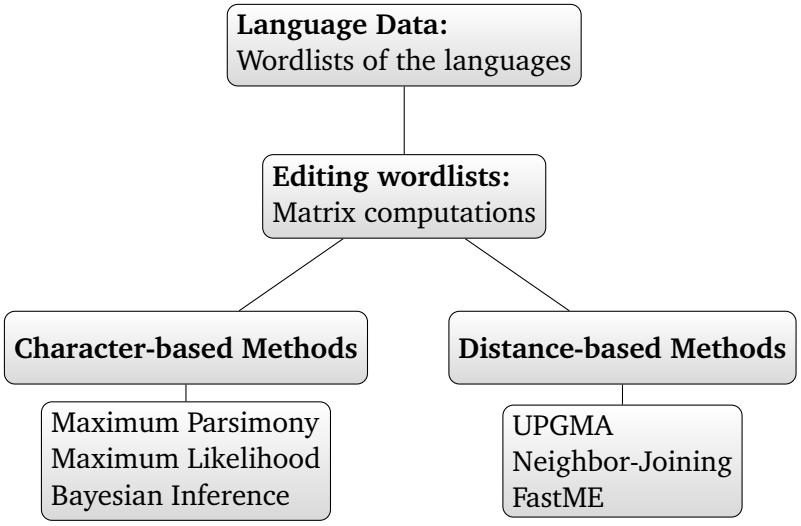


Fig. 3.1.: Workflow for automatic tree reconstruction.

For a better understanding of the various representations of phylogenetic trees, the properties, reconstruction methods, and labeling possibilities are explained in the following chapters. Depending on the data used for the reconstruction, the resulting trees present different information, and display various kinds of groupings of the entities under question, whereas the order of the processing steps remains the same. The distinct trees obtained from different kinds of linguistic data are the input to the HGT methods to detect horizontal transmissions. In addition, the computational steps, which need to be completed until the tree is compiled, are explained with respect to the process of the computational comparative method in CHL.

3.1.1 Trees as Graphs: The Fundamental Concept

From a mathematical point of view, the underlying concept for trees are graphs. Graphs are a popular way to represent the relationship between objects, where objects are represented as *nodes* and the relationships between them by *edges*. There are two types of graphs: *undirected* and *directed* graphs.[2] Huson et al. (2010, p. 3) define undirected graphs in the following way:

[1] For a detailed description and visualization of all alignment methods and tree reconstruction methods, see Huson et al. (2010, p.24).
[2] The overview is based on Huson et al. (2010).

Definition 3.1 (Undirected graph) *An undirected graph G = (V,E) consists of a finite set of nodes V and a finite set of edges E, where each edge $e \in E$ is of the form $\{v, w\}$, with $v, w \in V$.*

For an edge $e = \{v, w\}$, the endpoints of the edge are the nodes v and w, where e connects the two points with each other, i.e. v and w are incident to e. Two nodes are adjacent when they are connected by an edge, and two edges are adjacent when they share an endpoint. The degree of a node is the total number of incoming and outgoing edges (Huson et al., 2010). In an undirected graph, each edge consists of an unordered pair of two distinct nodes. In contrary, each edge in a directed graph consists of an ordered pair of nodes, i.e. the edge has a direction. Huson et al. (2010, p.4) give the following definition of a directed graph:

Definition 3.2 (Directed graph) *A directed graph G = (V,E) consists of a finite set of nodes V and a finite set of edges E, where each edge $e \in E$ is the form (v, w), with $v, w \in V$.*

In an edge $e = (v, w)$, v is the source and w is the target node connected by the edge e, the edge is directed from v to w. In this scenario, e is the out-edge of v and the in-edge of w. The indegree of a node is the number of in-edges, the outdegree of a node is the number of out-edges respectively. The degree of a node is therefore the sum of its indegree and outdegree. A directed graph which is free from directed cycles is called *a directed acyclic graph* (DAG). If there exists a direct path from node v to node w, v is the ancestor or parent of w, and w is the descendant or child of v. It follows that a node v cannot be both a descendant and an ancestor of any other node w (Huson et al., 2010). The definition of a DAG forms the basis for rooted phylogenetic networks, introduced later in the chapter.
A tree is a special kind of graph, i.e. a connected, noncyclic graph (Huson et al., 2010). A tree is transformed into a phylogenetic tree by labeling the leaves of the tree with the corresponding taxa contained in the set of taxa. The terminology introduced will be used to refer to the constituents of the tree as nodes and edges, whereas edges are sometimes called *branches* in phylogenetics (Huson et al., 2010). A phylogenetic tree can either be unrooted or rooted, where in a rooted tree one node is declared as root. Unrooted trees illustrate the relatedness of entities at the leaves without making any assumptions of ancestry. The declaration of a root results in a hierarchical classification of the data, representing the ancestry and the latest common ancestor of all entities at the leaves of the tree (Felsenstein, 2004; Huson et al., 2010; Morrison, 2011). This is in accordance with most of the manually compiled trees introduced so far. Biologists use rooted trees to display the evolution

of organisms, where each split in the tree is the result of a speciation event. In linguistics, rooted trees represent the ancestry of the languages and their splits into further unitary linguistic systems. This holds for both manually and automatically reconstructed trees.
Huson et al. (2010, p. 26) give the following mathematical definition of a rooted phylogenetic tree:

Definition 3.3 (Rooted phylogenetic tree) *Given a set of taxa χ, a rooted phylogenetic tree consists of a rooted tree $T = (V, E, \rho)$ and the taxon labeling $\lambda : \chi \to V$ that assigns exactly one taxon to every leaf and none to an internal node. All nodes, except ρ, must have degree $\neq 2$.*

In the definition, each taxon $x \in \chi$ is mapped onto some node v in the tree by the taxon labeling denoted as λ. In $\lambda(x) = v$, the node v is labeled by the taxon x. It is assumed that each leaf in the tree is obtained by some label via λ. In most cases, internal nodes remain unlabeled, and each leaf obtains exactly one label. Additionally, the definition requires that all nodes have a degree $\neq 2$.[3]
The set of taxa can be denoted as $\chi = \{x_1, ..., x_n\}$. In phylogenetics, each taxon x_i represents some species, group, or organism whose classification and evolutionary history is of interest for the analysis. In linguistics, the main aim of automatic tree reconstruction is the classification of languages according to the comparative method. The set of taxa consists of a set of languages, which are represented on the leaves of the tree. Whether other kinds of data can serve as input, and in which way the interpretation of the tree and its evolutionary history will differ, is one of the main topics in this thesis.
Definition 3.3 implies that any node in a tree can be specified as the root, i.e. inner nodes and leaves. Most, if not all of the tree reconstruction methods compile unrooted trees, since the selection of a root is an additional decision task in the computation. The decision task is either dependent on further knowledge of the evolution of the entities under question, or on mathematical assumptions. In classical historical linguistics, the tree is reconstructed bottom-up, answering the question of the root during the comparison and reconstruction process. The usage of computational methods leads to a rethinking in the selection of the best node to serve as root. There are several algorithmic solutions to this decision task, which are discussed later in the thesis. The mathematical definition of a phylogenetic tree opens up many possibilities to implement different algorithms for the reconstruction

[3] Mathematical phylogeny makes sometimes use of the more general concept of an X-tree (see e.g. Steel (2016)). For more information on this topic and an adequate discussion, see Huson et al. (2010, p. 25).

and manipulation of trees to address different research questions. Graph theory is widely used in different research fields where algorithms for manipulating and changing graphs can easily be adopted into phylogenetics. This results in a large number of algorithms which can be further integrated and adapted into CHL.
The rooted phylogenetic trees are of main interest to CHL, since they represent the evolution of languages in a hierarchical tree diagram and give a clear classification of the languages. The languages and their ancestry are of great interest in the detection of borrowing and loanwords. Borrowing can happen at any point in time in language history, which can be better indicated if the tree is hierarchically clustered. Additionally, rooted trees play an important part in the computation of rooted phylogenetic networks, which are important for displaying horizontal transmission due to the HGT methods, which can be used to detect borrowings and loanwords. From a theoretical point of view, rooted trees contain more information useful for further analyses and the detection of horizontal transmissions. In an algorithmic perspective there are other challenges to solve, which will be discussed below. Rooted trees are the trees of choice used in this thesis.[4]

3.1.2 Overview of Tree Reconstruction Methods

As it can be seen in figure 3.1, methods for phylogenetic inference are divided into two main methods: *distance-based methods* and *character-based methods*. The various algorithms are developed using different underlying mathematical models to reconstruct phylogenetic trees, which is one of the distinctions between them. However, another difference is the input matrices used for the reconstruction. The distance-based methods require a distance matrix, i.e. a matrix displaying the pairwise distances between two entities. Character-based methods require either a character matrix illustrating a multiple-sequence alignment, or a notation of the presence and absence of characters. The three main algorithms for each method are introduced below. For a detailed and mathematical description of the algorithms, please refer to Felsenstein (2004) and Huson et al. (2010), or to the original article for the corresponding method. An overview of the reconstruction algorithms in a linguistic context is given in Jäger and List (2016) and Dunn (2015b).

[4] For more detailed mathematical descriptions of directed graphs and trees in phylogenetics, see Huson et al. (2010). Delz (2013) gives an introduction to phylogenetic trees in a linguistic context.

Distance-Based Methods

In a distance-based analysis, the sequences are aligned using a pairwise alignment method while simultaneously computing a similarity or distance score. If the pairwise alignment results in a similarity score, the score is converted into a distance for further processing. The distance matrix is provided to an algorithm such as *UPGMA*, *Neighbor-Joining*, or *FastME* to reconstruct a pyholgenetic tree (Huson et al., 2010).

UPGMA The *unweighted pair group method using arithmetic means* (UPGMA) (Sokal, 1958) is a simple bottom-up hierarchical clustering method. UPGMA reconstructs a phylogenetic tree on a given distance matrix with edge length. "The method operates by clustering the given taxa, at each stage merging two clusters and at the same time creating a new node in the tree. The tree is assembled bottom-up, first clustering pairs of leaves, then pairs of clustered leaves, etc. Each node is given a height and the length of an edge is obtained as the difference of heights of its two end nodes" (Huson et al., 2010, p. 52). The algorithm presupposes a constant rate assumption, i.e. it assumes an ultrametric tree having equal branch lengths from the root to every node in the tree.

NJ The neighbor-joining algorithm (Saitou and Nei, 1987) is also a bottom-up hierarchical clustering algorithm. It reconstructs unrooted phylogenetic trees with branch lengths from a given distance matrix. "The neighbor-joining algorithm is a modification of the UPGMA algorithm. Both algorithms are agglomerative methods that repeatedly decide which two clusters to join, so that their nodes are 'neighbors' in the resulting phylogenetic tree" (Huson et al., 2010, p. 55). The UPGMA algorithm clusters correctly if the distances come from an ultrametric tree indicating that closest neighbors are indeed neighbors. However, in a general setting, two nodes or clusters can be separated by a small distance without being true neighbors in the tree. In this case, UPGMA reconstructs the wrong tree. To compensate for this effect, the NJ algorithm "subtracts the average distance (almost) of each cluster to all other clusters" (Huson et al., 2010, p. 54). NJ can therefore be used to reconstruct an unrooted phylogenetic tree from any distance matrix. In addition, the algorithm is faster and more widely applicable, which makes it the most popular method for computing trees in phylogenetics.

FastME The FastME algorithm (Desper and Gascuel, 2002; Desper and Gascuel, 2004) was developed within the framework of *balanced minimum evolution* (BME). It constructs an initial tree from a distance matrix using a distance-based algorithm, like NJ. Within the framework of BME, balanced average distances between two taxa are computed to assign a balanced edge length on the branch connecting the two taxa. The balanced length edge is defined depending on whether the edge connects an inner node and a leaf (leaf edge), or two inner nodes (inner edge). A mathematical description of the different formula is given in Huson et al. (2010, p. 57). The tree is then optimized using *nearest neighbor interchange* (NNI) and *subtree prune and regraft* (SPR), i.e. branch-swapping operations to transform a tree T into another T' by rearranging a part of T. NNI is a simple branch-swapping operation, in which subtrees that are attached to a common edge are swapped to find the optimal position (Moore et al., 1973; Robinson, 1971; Huson et al., 2010). SPR is a more general branch-swapping method, in which a subtree is pruned from the given tree T and regrafted at a different location of the tree (Page, 1993; Huson et al., 2010). The optimization steps are the reason why FastME runs significantly faster than NJ. FastME does not need to explicitly compute the edge lengths or tree lengths of the trees under consideration. The evaluation of the change of the tree length is associated with the number of insertions or swapping-moves, which can be computed in constant time. FastME has three advantages over NJ: it is faster, it provides more accurate trees, and it almost never produces negative branch lengths.[5]

Character-Based Methods

In a character-based analysis, there are three approaches to choose from: *Maximum Parsimony*, *Maximum Likelihood*, and *Bayesian Inference*. The input to these methods are character matrices, which are either obtained from multiple sequence alignment methods, or binary presence-absence data representations obtained from a clustering of the sequences, i.e. cognate clustering in linguistics. The optimal phylogenetic tree is usually found by the performance of a search in tree space, where it might be the case that several equally optimal trees are found. Character-based methods require an evolutionary model of the characters for the inference of the tree. The different models of evolution are described in Felsenstein (2004).[6]

[5] Huson et al. (2010, p. 37) give a detailed introduction to branch-swapping methods to solve optimization problems on phylogenetic trees.

[6] For an illustration of the classification of the most important models of DNA evolution, please see Huson et al. (2010, p. 31).

Maximum Parsimony The maximum parsimony method tries to find the optimal tree which explains a set of sequences by the minimum number of evolutionary events which happen during the evolution of the sequences along the tree. The maximum parsimony criterion is an optimality criterion, under which the tree that minimizes the number of character changes is preferred. The algorithm computes a parsimony score for each possible scenario of character evolution (Huson et al., 2010). The tree with the smallest parsimony score is the most optimal one. Since the algorithm reconstructs each possible scenario, it is possible that there exists more than one optimal tree with the same parsimony score, i.e. with the same number of character changes. It is therefore not trivial to choose the best tree if there is more than one optimal tree that fulfills the optimality criterion. The general idea of parsimony methods was first introduced by Edwards and Cavalli-Sforza (1963). The two most famous algorithms are introduced by Fitch and Margoliash (1967) and Sankoff (1975), and outlined by Felsenstein (2004).

Maximum Likelihood In phylogenetics, the basic idea of maximum likelihood estimation is to determine a phlyogenetic tree that maximizes the likelihood under a given model of evolution for the given data (Huson et al., 2010). The maximum likelihood approach uses standard statistical techniques to assign probabilities to a phylogenetic tree. The parameters of a probability function are computed by maximizing the likelihood function. The likelihood function is an indicator of goodness of fit for a statistical model to a dataset (Felsenstein, 2004). It is similar to maximum parsimony in the sense that the method requires a substitution model to assess the probability of mutations, and the more mutations a tree requires, the lower the probability. In other words, the less mutations needed to reconstruct a tree, the higher the probability. The tree with the highest probability is the most optimal tree obtained from the algorithm. The maximum likelihood approaches for phylogenies were introduced by Edwards (1964) using gene frequency data. Later, Felsenstein (1981) showed a practical computation for a moderate number of sequences using maximum likelihood estimation. An overview of maximum likelihood methods in phylogenetics is given in Felsenstein (2004), Salemi et al. (2009), and Huson et al. (2010).

Bayesian Inference The goal of Bayesian inference of phylogenetic trees is the estimation of the *posterior probability* of a pyhlogenetic tree. "Generally speaking, the *posterior probability* of a result is the conditional probability of the result being observed, computed *after* seeing a given input dataset. The *prior probability* of a result is the marginal probability of the result being observed, yet *before* the input

dataset is examined." (Huson et al., 2010, p. 45). In a Bayesian analysis, the dataset, the evolutionary model, and the prior probability of a tree need to be specified. The posterior probability of the tree can be obtained from the prior probability using *Bayes' theorem*. For more details on the computation, refer to Felsenstein (2004) or Huson et al. (2010). The computation of the posterior probability seems quite simple. However, it is usually impossible to solve the complete expression analytically. The computation of Bayes' theorem involves summing over all possible tree topologies and model parameters, which will be a time-consuming task, if it could be solved at all. To avoid this problem, the Bayesian inference algorithms use the *Markov Chain Monte Carlo* (MCMC) approach. The MCMC approach searches the tree space to find the region of the highest likelihood. A chain of trees is constructed, and at each step of the analysis a modification of the tree is proposed. The algorithm decides probabilistically if the modified tree is accepted or the current one is kept (Huson et al., 2010). Bayesian methods produce a tree sample according to the posterior distribution of phylogenetic trees instead of a tree representing a point estimate. This opens up new possibilities for further processing and analyses of the trees in the sample.

3.1.3 Language Trees: Genetic Classification of Languages

As described in section 2.1.1, genetic classification of languages aims at grouping languages according to shared innovations across a set of words. In accordance with the classical comparative method, the language relations are illustrated in a tree diagram to model diversity (Noonan, 2010; François, 2015). Linguistic phylogenetics incorporates the computational steps used in the computational comparative method to infer linguistic trees, illustrated in section 2.2.1 and visualized in figure 3.1. This chapter gives a short reiteration of the underlying process, and illustrates the computation steps to reconstruct linguistic trees displaying the genetic classification of the languages according to a basic vocabulary list.

To recap, phylogenetic trees are computed from sequence data, e.g. molecular data or language data. In linguistics, the single sequences are represented by words obtained from a basic meaning list of a set of languages. The total number of languages and meanings is restricted by the database used for the computation. The size of the meaning list varies across databases, e.g. between 40 words in ASJP, and 1,016 word in NELex. The introduced databases provide basic meaning lists containing phonological representations of the words, which is important for the detection of sound correspondences. A word list, as shown in table 3.1, serves as starting point for the analysis. For the reconstruction of language trees, the whole

basic meaning list is used as input for all computation steps. The cognate assignment is not available in all databases, and is therefore assigned manually for the following presentation of the example.

	Meaning 'sun'		Meaning 'mountain'		Meaning 'you'		Meaning 'louse'		
	word	cc	word	cc	word	cc	word	cc	
English	son	s	mauntin	m	yu	d	laus	l	...
Icelandic	soul	s	fEtl	f	8u	d	lus	l	...
Swedish	sul	s	bErg, fyEl	b,f	d3	d	l3s	l	...
Norwegian	sul	s	bErg, fyEl	b,f	d3	d	l3s	l	...
Dutch	zon	s	bErx	b	y3	d	l3is	l	...
German	zon3	s	bErk	b	du	d	laus	l	...
Romanian	soare	s	munte	m	tu	d	p3duke	p	...
Italian	sole	s	monta5a	m	tu	d	pidokyo	p	...
French	solEy	s	mota5	m	ti	d	pu	p	...
Catalan	soL	s	munta53	m	tu	d	pol	p	...
Portuguese	soL	s	mota5a	m	tu	d	pyulu	p	...
Spanish	sol	s	monta5a	m	tu	d	pyoxo	p	...
...									

Tab. 3.1.: An excerpt of the data from the NorthEuraLex Database containing the phonological representations of the words in the ASJP alphabet and manually assigned cognate classes.

Alignment analyses are used for sequence comparisons and detection of sound correspondences across all languages in the set. No previous knowledge or expert information of language relationships is needed for the alignment. For each meaning in the list, all languages are compared and aligned with one another. Pairwise sequence alignment methods simultaneously compute a similarity or a distance score for each alignment, whereas a similarity score is converted into a distance for further processing. At the end of the computation, the distance scores are aggregated to maintain an overall distance between two languages across all meanings. The pairwise distances between the languages are visualized in a matrix, where the distance scores are the off-diagonals and the diagonal is filled with 0, indicating the relation between the same language. Table 3.2 shows a sketch of a distance matrix for the Germanic languages, using arbitrary distance scores for illustration purposes.

	eng	isl	swe	nor	nld	deu
eng	0.0	0.36	0.38	0.38	0.3	0.3
isl	0.36	0.0	0.1	0.1	0.2	0.2
swe	0.38	0.1	0.0	0.08	0.25	0.24
nor	0.38	0.1	0.08	0.0	0.24	0.25
nld	0.3	0.2	0.25	0.24	0.0	0.01
deu	0.3	0.2	0.24	0.25	0.1	0.0

Tab. 3.2.: A sketch of a distance matrix for the Germanic languages using arbitrary distance scores.

The distance matrix obtained from the pairwise sequence alignment serves as input for all distance-based methods. In linguistics, the neighbor-joining algorithm and FastME are the most popular ones (Jäger, 2013a). UPGMA is suited best for ultrametric data, since sequences are in most cases not ultrametric, neighbor-joining and FastME are the better choice for linguistic data. In addition, NJ and FastME are way faster and produce more correct trees than UPGMA (Huson et al., 2010).
The distance scores and sound correspondences received from the pairwise alignment serve as indicators to determine cognate classes for the corresponding meanings. Cognates are either assigned automatically, or manually by linguists. There are various methods for automatic cognate detection which are based on sequence alignments and their corresponding distance scores. For each meaning, depending on the alignments and distances of the words, the algorithm assigns cognate classes to the words, and simultaneously clusters them into cognate sets. The result is a categorization, as it can be seen in table 3.1. The cognate classes are represented in a presence-absence character matrix, like in table 3.3, where 1 means presence and 0 absence of the cognate class.

	Cognate sets							
	s	m	b	f	d	l	p	...
English	1	1	0	0	1	1	0	...
Icelandic	1	0	0	1	1	1	0	...
Swedish	1	0	1	1	1	1	0	...
Norwegian	1	0	1	1	1	1	0	...
Dutch	1	0	1	0	1	1	0	...
German	1	0	1	0	1	1	0	...
Romanian	1	1	0	0	1	0	1	...
Italian	1	1	0	0	1	0	1	...
French	1	1	0	0	1	0	1	...
Catalan	1	1	0	0	1	0	1	...
Portuguese	1	1	0	0	1	0	1	...
Spanish	1	1	0	0	1	0	1	...
...								

Tab. 3.3.: The presence-absence matrix for the languages and cognate classes displayed in table 3.1.

The character matrix in table 3.3 is used as input for the character-based methods. Depending on the method of choice, the result will either be a tree representing a point estimate, as for maximum parsimony and maximum likelihood approaches, or a distribution of trees from Bayesian models.

The language tree can be compared to a species tree in phylogenetics, i.e. a tree displaying the evolution and relatedness of species or organisms. This representation is the most popular one in both fields, and the algorithms are mainly developed for this reconstruction purpose. In contrast to phylogenetics, automatically reconstructed language trees are of main interest in CHL so far. There are two reasons: first, one of

Fig. 3.2.: The distance-based language tree, representing the language classification of the selection of Germanic and Romance languages.

the main aims in CHL is the implementation of the comparative method to classify languages automatically, and second, different databases provide various kinds of input data needed for the automatic reconstruction of languages. In addition, there are several databases like *Glottolog* (Hammarström et al., 2018) providing expert classifications of linguistic trees, which can be used to evaluate the automatically reconstructed trees obtained from the algorithms.

The language tree in figure 3.2 is a distance-based bifurcating tree computed using FastME, which is the output of most tree reconstruction methods, except for Bayesian inference. The Bayesian analysis infers a distribution over trees which could be summarized into a single consensus tree using different algorithms.[7] Linguists debate whether bifurcating or multifurcating trees are best to represent language-relatedness. A language can have more than two descendants, since the splitting of the languages or varieties into unitary systems is not restricted to two languages per generation. This is a valid assumption, which is reflected in the expert classification provided by Glottolog. Different linguistic studies showed that a bifurcating tree obtained from tree reconstruction methods corresponds to over 80 percent to the expert tree from Glottolog, see e.g. Steiner et al. (2011), Jäger (2013a), and Jäger (2013b). Those results show that bifurcating language trees are suitable for representing language classifications.

[7] The consensus tree methodology is introduced in chapter 4.

In a language tree, adjacent languages share more sound correspondences than less related languages. This is reflected in the clustering of the tree in figure 3.2. The classification of the languages is in accordance with the linguistic knowledge: all Germanic languages and all Romance languages are clustered together. Taking a look at the detailed clustering of the Germanic languages, the clusters correspond to the linguistic and geographical background knowledge, i.e. German and Dutch are closely related and clustered together, the same holds for Swedish and Norwegian, including Icelandic. McMahon (2010) showed that there are a considerable number of loanwords within the basic meaning list, proving that loanwords and borrowings cannot be ignored in the analysis. However, as it was shown in section 2.1.3, the loanwords do not disturb the signal for an appropriate clustering of the languages, which is reflected in the automatically reconstructed tree. English is still within the cluster of the Germanic languages, since the majority of the words in the basic meaning list are cognates to other Germanic languages. Nevertheless, the loanwords (explicitly the French loanwords) are responsible for grouping English as outlier of the Germanic languages in the sample.
To sum up, an automatically reconstructed language tree reflects the language relationships in correspondence to the classical comparative method, illustrating the language families and splits into unitary linguistic systems. In the hierarchical representation of the tree, two nodes (either leaves or inner nodes) are joined together, where the mother node represents the most recent common ancestor. The language tree represents the relatedness of the languages according to their clustering into language families.

3.1.4 Concept Trees: Genetic Classification of Words

It has been known for a long time in historical linguistics that words can evolve differently and independently of their languages. The famous statement "every word has its own history"[8] is often credited to Gilliéron and Edmont (1902-1910), although Campbell (2013) attributes it to Hugo Schuchardt, one of the founders of the wave model. The assumption that every word has its own history is the fundamental idea of the wave theory to model the spread of words in accordance to their innovations with regard to the genetic classification represented in language trees.
Words can enter a language either through vertical or through horizontal transmission, i.e. they can be inherited from the ancestor language, transferred due to language contact, or derived from other words. All the words belonging to the

[8]Translated from the original: "chaque mot a son histoire" (Gilliéron and Edmont, 1902-1910).

language, independent of their transmission, can evolve and change in different ways during language evolution. Words can change their form by sound changes and analogy, they can change their meaning due to semantic change, or they can disappear completely from the language. The result is an individual development of the words during the evolution of the language.

Within the framework of the classical comparative method, it was already known that words within a language might evolve differently than others. The distinction of different sound changes, the assignment of shared innovations to a specific word group, and the detection of complementary distributions within correspondence sets are the main indicators for language classification and the reconstruction of proto-languages.[9] Other mechanisms, like analogy, semantic change, and extinction, play also a major role in the evolution of languages, although they receive less attention within the comparative method. Regarding the process of horizontal transmission, Haugen (1950) classified the different kinds of loanwords by dividing them into categories with respect to their adaptation and change during the borrowing process. The adaptation process is an individual and language-dependent procedure. Regarding the language and the phonology of the borrowed word, the word is changed to fit into the structure and phonology of the RL. After the integration of the loanwords, the words can undergo the same evolutionary changes as other words in the language, i.e. they are treated like inherited words. The different processes of evolution and change of words shape the characteristics of a language, and are responsible for language diversity.

The evolution and classification of languages according to their shared innovations across a list of basic meanings is displayed in a language tree. This representation is not sufficient enough to illustrate the individual evolution of words. The genetic classification of words aims at grouping the words within a single meaning according to their shared innovations. The language clusters represent the classification of languages according to one single concept. The resulting tree is called a *concept tree*.[10]

Phylogenetic trees are computed out of sequence data, which could also be a sequence of sounds in a word. Concept trees represent the clustering of the languages according to a single meaning in the basic vocabulary list, illustrated in table 3.4. The number of languages is therefore restricted by the database used for the analysis. The number of words determines the number of resulting trees of the analysis, i.e. the evolution of each meaning is represented by a single tree. The size of the

[9] A summary of different kinds and processes of sound changes can be found in Campbell (2013).

[10] The word "concept" and "meaning" can be used interchangeably. Here, the term "concept" is used to refer to one meaning in the basic meaning list. Since the word for a meaning in the list can consist of more than one part, the term "concept" is preferred over the term "word".

meaning list is also restricted by the database, since each database provides its own meaning list.

	Meaning 'mountain'	
	word	cc
English	mauntin	m
Icelandic	fEtl	f
Swedish	bErg, fyEl	b,f
Norwegian	bErg, fyEl	b,f
Dutch	bErx	b
German	bErk	b
Romanian	munte	m
Italian	monta5a	m
French	mo,mota5	m
Catalan	munta53	m
Portuguese	mot3, mota5a	m
Spanish	monta5a	m
...		

Tab. 3.4.: An excerpt of the data from the NorthEuraLex Database containing the phonological representations of the concept 'mountain' in the ASJP alphabet and manually assigned cognate classes.

The computation of the matrices and the reconstruction of the trees are similar to the ones for language trees. Alignment analyses are used for sequence comparisons and for the detection of sound correspondences between the words. Simultaneously, the pairwise alignment methods compute a similarity or distance score, where a similarity score is converted into a distance for the matrix generation. The pairwise word distances are visualized in a language matrix, where the diagonals are filled with 0 and the off-diagonals with the pairwise distances between the words of two languages. A sketch of a distance matrix for the concept 'mountain' using a small language sample is given in table 3.5.

	fra	spa	eng	deu	nld
fra	0.0	0.2	0.3	0.98	0.98
spa	0.2	0.0	0.33	0.95	0.95
eng	0.3	0.33	0.0	0.93	0.93
deu	0.98	0.95	0.93	0.0	0.05
nld	0.98	0.95	0.9	0.05	0.0

Tab. 3.5.: A sketch of the concept distance matrix for 'mountain' of a language sample using arbitrary distance scores.

The language sample is chosen in order to illustrate the small distance between the English word *mountain* and its counterparts in the Romance languages. The small distance indicates that the English word shows a greater similarity to the words of the Romance languages compared to the ones present in the Germanic languages, i.e.

the English word *mountain* is a loanword. The distance matrix serves as input for the neighbor-joining or FastME algorithms to reconstruct the concept trees. Compared to the matrix computation for language trees, the distance matrices for concepts can capture the history of the words in more detail and give a better insight into the language relationships within one concept. It can be assumed that the language relationships according to the word pair distances are reflected in the reconstructed concept trees.

The manually or automatically detected cognate classes can be used to reconstruct a binary character matrix, which serves as input for the character-based tree reconstruction algorithms within the frameworks of Maximum Likelihood or Bayesian inference. The generation of a binary presence-absence matrix of single meanings can lead to one difficulty, namely the presence of too few cognate classes. Few or even one cognate class illustrate the relatedness of all words from a single common ancestor. Although this clustering is highly informative for linguists, the resulting character matrix would contain too few data points for a character-based analysis. The reconstruction of a tree with enough phylogenetic signal using character-based methods is therefore questionable. New ideas to gain more signal out of the characters of the words to reconstruct suitable character matrices for the tree reconstructions are introduced in chapter 4.2.

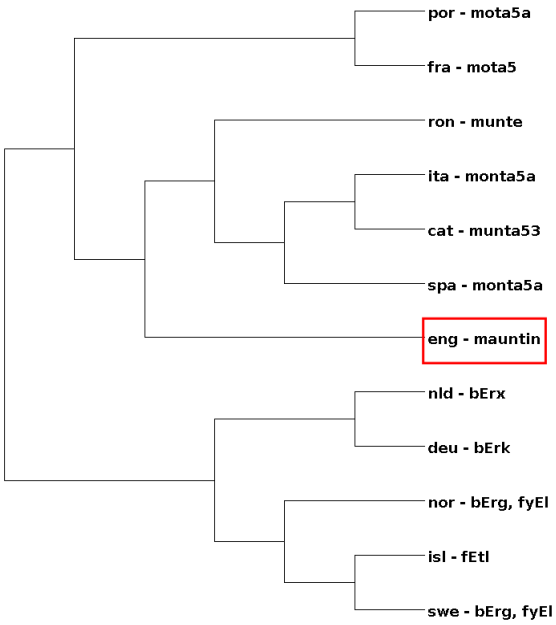


Fig. 3.3.: The concept tree 'mountain' representing the language classifications of the selection of Germanic and Romance languages, reconstructed using a distance-based method.

The concept tree can be compared to a gene tree in phylogenetics. Gene trees represent the evolution of individual genes, which is said to be different to the evolution of species. In biology, a mixture of the genes is transferred from the parents to the offspring due to recombination. Some organisms can even transfer genes without sexual reproduction. Similar to words, genes can be inherited, transferred, and changed during time according to different mutation processes. In phylogenetics, the research focuses on the reconstruction of gene trees as input to further analyses and on the fusion of gene trees to obtain a species tree, where the species tree is unknown or not reconstructable. In linguistics, algorithms for reconstructing and summarizing concept trees can be adapted and used to get insights of the evolution of words.

Figure 3.3 displays a distance-based concept tree for the meaning 'mountain' obtained from FastME. Similar to the language tree, the concept tree is a bifurcating tree. As already stated, bifurcating trees are suitable to represent language classifications. With respect to further processing and the reconciliation of the trees to detect horizontal transmission, it is essential that both trees share the same branching pattern.

A concept tree reflects the language classifications according to a single meaning in the list. The Germanic and Romance languages are still separated, forming their own language clusters as a family. English, however, belongs to the cluster of the Romance languages, which is clearly reflected in the concept tree. Historical linguistic research has shown that the English word *mountain* was borrowed from Old French (Merriam-Webster Dictionary, 2020). Since the sequence comparison is done using phonological representations of the modern languages at this point in time, the clustering of the languages in the concept tree might change due to the evolution of the words. The languages are clustered according to their alignment and distance scores, without any insight into their history or older word forms.

This is a shortcoming of the usage of phylogenetic reconstruction methods, since explicit processes like sound change, word formation, inflection, analogy, compounding, etc. are not modeled and reconstructed in the tree. List and Schweikhard (2020) proposed a framework of modeling individual word histories, including relevant processes of word changes. The focus on this study lies in modeling the formation of words throughout history, taking regular sound changes, derivations, borrowing, and other processes into account which could have influenced the evolution of the words. However, the model requires an annotated database, including roots and attested forms. List and Schweikhard (2020) manually selected and annotated the data from the etymological dictionary *Nomina im indogermanischen Lexikon* (NIL) (Wodtko et al., 2008), which the reconstruction of word trees relies on. Publicly available databases include basic meaning lists of varying sizes without further

annotations or reconstructions of the word histories. The extension of databases like NELex with these kinds of annotations and assignments as proposed by List and Schweikhard (2020) is time-consuming, if possible at all. In their framework, List and Schweikhard (2020) used the selected and annotated data from Proto-Indo-European languages, one of the best-attested language families so far. For other languages and language families which do not have this amount of written records and research on the reconstruction of the languages, this model could not be applied. If the model is fully developed and transferable to other databases, the word trees would be a great achievement in computational historical linguistics, and could be used for further analyses in terms of tree reconciliation.

In addition, List and Schweikhard (2020) mentioned that the comparison of all words in one concept using alignment methods indicates that all words are related to each other, which seems to be unrealistic in a historical linguistics context. However, the further back in time words and their forms are reconstructed, the less information about word forms and their meaning are present. The question is whether there is enough certainty if the reconstructed form is clearly the most latest common ancestor of the current occurrence of the word. Needless to say, there is a lot of substantive and verified research in historical linguistics and the reconstruction of proto-languages with respect to the comparative method. However, Kessler (2001) stated with regard to loanwords that most of the time the certainty whether a word is a loanword fades at some point in the reconstruction, which makes a clear distinction and clarification of the word history nearly impossible. Unfortunately, the information on word histories from etymological dictionaries, including a detailed description of the cognate and loanword status, is not included in any of the publicly available databases. The question whether all words and all languages descend from one single ancestor cannot be answered with clarity and remains open.

In this thesis, the focus lies on the reconstruction of concept trees using standard matrix computations and tree reconstruction algorithms. To my knowledge, there is no other study testing different matrix computation approaches with respect to concept tree reconstructions using linguistic data, which can further be used for tree reconciliation algorithms to detect horizontal transmission. In chapter 4, various methods on matrix generations are introduced with respect to linguistic data on basic vocabulary lists.

Distinction between Language and Concept Trees

The evolution of words can be independent and different to the evolution of their corresponding language. This leads to the idea to reconstruct language trees and

concept trees for the comparison of the language classifications in order to obtain insight in different evolutionary processes like borrowing.
Language trees represent the history and evolution of languages across a list of basic meanings. The clustering illustrates the relatedness of the languages according to shared innovations over a set of words. This indicates the overall relationship of languages, their ancestry, and evolution. Concept trees, on the other hand, illustrate the history and evolution of single concepts and their phonological representations in the languages under consideration. The clustering displays the grouping of the languages according to shared innovations within one single meaning. This gives new insights in the relation of the languages conforming to a single concept, which opens up the view on word-relatedness due to inheritance and language contact.
The mathematical concept of a phylogenetic tree is the fundamental concept of both trees. Additionally, the tree reconstruction methods provided by phylogenetics are used to reconstruct both types of trees. The main difference is the input data and the representation purpose. The language tree is reconstructed using all words in the basic meaning list, resulting in one single tree representing the language classifications. The concept tree is reconstructed out of all words for a single meaning in the list, where the number of resulting trees is determined by the number of meanings in the list. Each single tree illustrates the language classification according to a meaning in the list. The history of the words can differ, and so does the clustering of the languages on the concept tree. The trees are used for further processing in order to detect borrowings, and especially loanwords. Since the clustering of the languages might differ in the language and the concept tree, the trees can be compared to detect matches and mismatches in the branching patterns of the trees. For these kinds of algorithms, it is important that the underlying tree concept and the branching pattern are similar in both trees, i.e. both trees are in accordance with the mathematical definition of a phylogenetic tree. In order to fulfill these requirements, similar reconstruction processes are chosen for language and concept trees. Since concept trees are not well researched and bear different challenges, different matrix reconstruction methods are introduced in chapter 4.

3.2 Phylogenetic Networks in Linguistics

The concept of the tree is the fundamental principle in both phylogenetic and linguistic research. Groundbreaking research has shown the importance of the concept to model the classification of species and languages, or of genes and words, and analyze the results statistically. However, the significance of horizontal transmission within

the evolutionary process was long underestimated and paid less attention to in the classification task. It is long known in historical linguistics that language contact cannot be dismissed as trivial in the evolution of languages. Not even the basic vocabulary is resistant to borrowing, which underlines the importance of including language contact and horizontal transfer into the analysis of language evolution.
Phylogenetics provides network models as an alternative approach to phylogenetic trees, which allow to display horizontal transmission in the evolutionary process. Still, the tree concept serves as the underlying model for most networks models, i.e. the fundamental principle is to reconstruct a tree in cases where the data contains no amount of horizontal transmission, and to reconstruct a network for datasets with horizontal signals. Networks are therefore suitable for taxa that evolve in a tree-like manner, i.e. a small amount of horizontal transmission is assumed for the dataset, as well as for datasets including a high amount of horizontal transmission, i.e. reticulate events (Huson et al., 2010). Reticulate events in phylogenetics are caused by processes like horizontal gene transfer, hybridization, recombination, and other processes. In linguistics, reticulate events are caused by language contact and the transfer of linguistic items of the languages. In addition, networks can represent conflicts in the data that are, for example, due to mechanisms like incomplete lineage sorting, or to deficiencies of the assumed evolutionary model (Huson et al., 2010). Like trees, networks can have different graph theoretic properties and representation purposes. The properties of the networks exemplify the type of networks, which at the same time describe and visualize specific evolutionary scenarios in the phylogeny. Depending on the linguistic dataset and the research question, the adequate network model needs to be chosen. Not all types of networks are suitable to display horizontal gene transfer or the transfer between languages.
For a complete understanding of the network models, the different network types and their properties are introduced in this chapter. In addition, several kinds of networks are compared to each other in order to find the most suitable one to represent horizontal transmission with respect to the linguistic requirements of borrowing and loanword detection. Finally, an overview of different linguistic studies using networks to model language contact and reticulations are introduced and discussed.

3.2.1 Networks: Underlying Concept and Application

The envisioned idea of phylogenetic networks in biology is to elaborate the details of a (rooted) phylogenetic network to describe the evolution of life, as it was proposed by Doolittle (1999). Ideally, the networks represent the history of organisms,

taking reticulate events into account. However, similar to linguistics, horizontal transmission can be caused by different mechanisms. In phylogenetics, different networks and different analyses are proposed to describe specific mechanisms and processes of horizontal transmissions. This results in approximately 20 different names and definitions for networks proposed in the phylogenetic literature (Huson et al., 2010).[11] A more general distinction between networks can be made according to the categories, depending on their graph theoretic properties and representation purposes, as shown in figure 3.4.

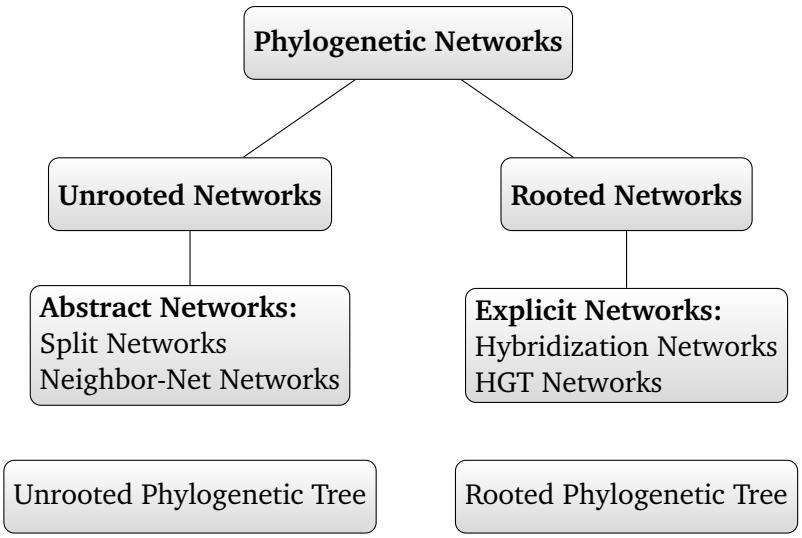


Fig. 3.4.: Overview of networks and algorithms important for this thesis.

From a mathematical point of view, the underlying concept of a phylogenetic network is an (un)directed cyclic graph, which is a generalization of an (un)directed acyclic graph, as defined in section 3.1.1. In networks, objects or taxa are represented as nodes, the relationship between the objects by tree edges, and horizontal transmission by *reticulations*, i.e. *reticulate edges*.[12] Networks can be either unrooted or rooted, where rooted networks are used to illustrate evolutionary history, including horizontal transmission. A second major distinction between networks is the one between *data-display* and *evolutionary* networks, also named *abstract* and *explicit* networks respectively. Data-display networks visualize incompatible datasets with an amount of horizontal transmission in a fruitful manner, where conflicts in the dataset can occur due to mechanisms such as incomplete lineage sorting. Evolutionary networks, on the other hand, illustrate a putative evolutionary history involving horizontal transmission, i.e. reticulate events caused by hybridization

[11] The process of naming networks according to specific properties is clarified in Huson et al. (2010, p. 71).

[12] For a graphical illustration and description of the terminology, see Morrison (2011, p. 41–42).

or horizontal gene transfer. By definition, most unrooted networks are abstract networks, and thus, an unrooted tree is also abstract in the sense of displaying the characteristics of the dataset. Rooted phylogenetic networks can be either abstract or explicit, depending on the underlying technique of their reconstruction algorithm (Huson et al., 2010).
Huson et al. (2010, p. 71) give a very general definition of an unrooted network.

Definition 3.4 (Unrooted phylogenetic network) *Let χ be a set of taxa. An unrooted phylogenetic network N on χ is any unrooted graph whose leaves are bijectively labeled by the taxa in χ.*

In phylogenetics, each kind of unrooted network has another underlying method, which leads to an extension and specification of the mathematical definition. Huson et al. (2010) focus on the most prominent networks used in phylogenetics, the class of *split networks* and *quasi-median networks*.[13] Phylogenetics provides a great number of established tools for the reconstruction of unrooted phylogenetic networks, which can be directly adapted into CHL and applied to linguistic data. It is therefore not surprising that the first studies in CHL applied methods to reconstruct unrooted networks to display language contact. Keeping in mind that abstract networks display the discrepancies in the dataset, the reticulations can be interpreted as an occurrence due to horizontal transmission. However, unrooted networks display the relationship between the languages without making any assumptions of ancestry. The evolutionary history of the languages can therefore not be reconstructed in terms of inheritance.[14]
In the phylogenetic analysis of reconstructing rooted trees and networks, the concept of clusters plays an important role. For a set of taxa, clusters represent a group of elements which share an attribute. In linguistics, clusters represent languages according to shared innovations, like sound correspondences and cognate classes. A rooted phylogenetic tree represents a set of *compatible* clusters, i.e. languages are grouped according to their shared innovations to form language groups and families. A rooted phylogenetic network additionally represents *incompatible* clusters, which may occur due to evolutionary events like horizontal gene transfer in biology and borrowing in linguistics (Huson et al., 2010). Since the underlying graphical structure is a DAG, rooted phylogenetic networks are a generalization of rooted phylogenetic trees. This is reflected in the definition of rooted phylogenetic networks given by Huson et al. (2010, p. 138):

[13] The mathematical and algorithmic descriptions are described in Huson et al. (2010).
[14] See Morrison (2011) for a detailed description of data-display networks.

Definition 3.5 (Rooted phylogenetic network) *Let χ be set of taxa. A rooted phylogenetic network $N = (V, E, \lambda)$ on χ consists of a directed graph $G = (V, E)$ and a node labeling $\lambda : \chi \to V$ such that:*

(i) The graph G is a directed, acyclic graph (DAG).
(ii) The graph G is rooted, that is, has precisely one node ρ of indegree 0.
(iii) The node labeling λ assigns a taxon to each leaf of G.

We usually assume that the network does not contain any node that is suppressible, in other words, any unlabeled node with indegree 1 and outdegree 1. In addition, we usually assume that the taxon labeling λ is a bijection between the taxon set χ and the set of leaves of N.

The definition is specified in a general way, since rooted phylogenetic networks can be restricted in various ways, depending on the evolutionary processes under consideration. In the case of horizontal gene transfer, it is assumed that only two species can be involved in a transfer event. The set of indegrees of nodes is restricted with respect to the process of HGT, i.e. the values are determined to 0, 1, and 2. The root of the network has an indegree of 0, tree nodes have an indegree of 1, and reticulate nodes an indegree of 2, respectively (Huson et al., 2010).[15]
At this point, it is important to mention that in the case of phylogenetic trees, unrooted trees can be rooted by choosing one node in the tree as root node, i.e. the tree can be rooted using any internal node. The resulting rooted tree can be interpreted as displaying the evolutionary history of the entities. This means all tree reconstruction methods independent of the resulting trees can be used for building a rooted phylogenetic tree and making assumptions about the ancestry of the languages. This is, however, not the case for phylogenetic networks. Unrooted networks cannot be rooted to display the evolutionary and horizontal relationships between the objects. A data-display network is an interpretation of the data in terms of interconnections. The nodes in a data-display network play a heuristic role, displaying the conflict of the data, and the edges simply represent this difference between the nodes (Morrison, 2011). If a data-display network is rooted, the result would be a nonsensical diagram where the nodes do not represent the ancestors of two other nodes, and the edges do not represent horizontal transmission due to transfer processes. Rooted, and especially evolutionary, networks need to be computed using different algorithms in order to display the evolutionary relationship

[15] The different types of nodes and edges present in a rooted phylogenetic network are listed in Huson et al. (2010, p. 139). Additional terminology for networks is explained in Huson et al. (2010, p. 140).

between ancestors and descendants in terms of inheritance and horizontal transmission.

The underlying assumptions on evolutionary networks to model horizontal gene transfer are in accordance with the fundamental principles of borrowing in historical linguistics. The process of borrowing involves only two languages which are in contact with each other. A node in the network with indegree 2 indicates two evolutionary processes, one on the basis of inheritance, displayed by the edge connecting the mother with the daughter language, and one process due to horizontal transfer, connecting the two languages which are in contact with each other. Depending on the network of choice, the two edges can clearly be identified and assigned to the corresponding processes. This is not the case for all types of evolutionary networks, which means that the method and algorithm for reconstructing evolutionary networks should be chosen carefully depending on the aim of the analysis and the research question. In computational historical linguistics, evolutionary networks are rarely used. One of the main reason is that the calculation of rooted phylogenetic networks is difficult and fewer computational methods are developed for practical usage. Within the last decade, the development of reconstruction algorithms for evolutionary networks emerged in the field of phlyogenetics. The adaptation and application of these methods in CHL is still in its infancy. However, to model the process of borrowing and to detect loanwords, the focus lies on the application and adaptation of HGT algorithms. These algorithms generate evolutionary networks where inheritance and horizontal transmission can clearly be identified. An introduction to the fundamental idea of horizontal gene transfer and the corresponding network model is given in chapter 3.3.

3.2.2 Network Analyses in Linguistics

Data-display networks, especially neighbor-net networks, and some evolutionary networks are used in linguistic studies to model the connections between languages due to inheritance and contact. All studies operate on the language level to shed light on the complete language evolution of the Indo-European language family. However, some methods introduced in the studies are used to address research questions in other language families as well. Not all networks used in linguistic studies focus on the detection of borrowing and loanwords, but aim at modeling various kinds of secondary connections due to horizontal transmission. In this section, the usage and the main aims of the different kinds of networks used in linguistic studies are introduced and discussed. Although most of those networks and their algorithms

cannot be used to explicitly detect borrowing and loanwords, they lead to a better understanding of the evolution of natural languages.

Data-Display Networks

Linguistic studies using split networks to visualize language relationships are interested in the so-called secondary connections between languages. Split networks can be reconstructed using various types of linguistic data, including lexical data (Hamed and Wang, 2006; Bryant et al., 2005), phonetic data (McMahon et al., 2007; Heggarty et al., 2010; Prokic, 2010), and typological data (Daval-Markussen and Bakker, 2011; Szeto et al., 2018). While some studies focus on representing all kinds of historical relationships within a language family (Lehtinen et al., 2014; Heggarty et al., 2010), others aim at detecting contact and borrowing between the languages (McMahon and McMahon, 2005).

Kessler (2001) mentioned that at some point in prehistory, the question whether a word is a loanword becomes unanswerable. The first major distinction is therefore the one between any kind of historical connections and similarities due to chance. Languages are historically connected if they share certain properties, no matter if the elements are inherited from a common ancestor or transferred via language contact. Shared elements due to horizontal transmission cause noise in the tree model, which is not reflected in the representation. Data-display networks are an adequate combination of the tree and the wave model, while aiming at visualizing various kinds of historical relations between the languages. The noise in the dataset can be identified by discrepancies, which lead to incompatible evolutionary scenarios displayed as reticulate edges. One of the most prominent methods to reconstruct split networks is the neighbor-net algorithm, introduced by Bryant and Moulton (2004). The algorithm uses a similar idea to reconstruct networks than the neighbor-joining algorithm to reconstruct phylogenetic trees. A distance matrix is generated from the dataset, which serves as input to the algorithm to cluster the languages according to shared innovations. The results of the clustering algorithm can lead to overlapping clusters, which are then visualized in an unrooted network.[16] The studies using split networks aim at modeling the complete evolutionary history of the languages under question, taking common ancestry and horizontal transmission into account (McMahon and McMahon, 2005; Heggarty et al., 2010; Lehtinen et al., 2014). The detection of borrowing events, loanwords, and the direction of the

[16] For a detailed description of the neighbor-net method, see Bryant and Moulton (2004) or Huson et al. (2010, p. 254). The algorithm is implemented in the SplitsTree software by Huson and Bryant (2006) and in the R package *phangorn* (Schliep et al., 2017).

transfer is disregarded. Since the underlying evolutionary processes causing the scenarios cannot be captured by a split network, it is also questionable if and how horizontal connections can be distinguished from similarities by chance.[17]

In their study, McMahon and McMahon (2005) claimed that borrowing must be the prime candidate for the reticulations they hope to detect by means of discrepancies in the dataset. A character-based algorithm called *Network* (Bandelt et al., 1999; Forster et al., 2001) is used to reconstruct networks on real language and simulated data.[18] The program results in a split network, where a dataset without discrepancies is represented as a tree and otherwise as a network. The reticulations in the network should represent language contact due to borrowing. However, the network cannot display any further information on the direction of borrowing or on the items responsible for the reticulations, i.e. the items which are detected as loanwords by the algorithm. There are two different ways to test which items are actually causing the discrepancies in the data. One way would be sampling the data to generate different networks and evaluating them by counting the emergence of the same (sub)patterns. This process is called either *bootstrapping* or *jackknifing* in phylogenetics, depending on the sampling technique. The sampling processes, depending on the size of the data and the number of samples, are time-consuming tasks.[19] Since the network algorithm provides a shortcut, McMahon and McMahon (2005) discarded the sampling procedure. For each graph produced, the algorithm provides a list of data points which were difficult to reconcile with the tree, i.e. the items which cause the algorithm to insert reticulations. However, the number of reticulations, and therefore also the list of data points, depend on an internal parameter, which needs to be set manually. The parameter reflects the sensitivity of the algorithm according to the visualization of incompatible data structures. Parameter setting is always a challenging task. The best parameter can only be found by evaluating the results, which would lead to a circular process. The user needs to estimate which amount of reticulations, i.e. which amount of loanwords, would be expected in the given dataset. This requires linguistic knowledge of the languages in the dataset and an expectation of the amount of loanwords and language contact between the languages. Otherwise, the parameter needs to be set arbitrarily. McMahon and McMahon (2005) discussed the problem in their study, but did not provide a solution for the parameter setting. The results of their study showed that for a small Germanic/Romance dataset, Network detects the expected reticulations.

[17] A graphical representation of the networks modeling secondary connections can be found e.g. in Lehtinen et al. (2014), Heggarty et al. (2010), and McMahon and McMahon (2005).

[18] The Network program (Bandelt et al., 1999; Forster et al., 2001; Polzin and Daneshmand, 2003) can be found online under `https://www.fluxus-engineering.com/sharepub.htm#a6`. For further information on the application in linguistics, see McMahon and McMahon (2005) and McMahon et al. (2005).

[19] A detailed description of the processes is given in chapter 4.

McMahon and McMahon (2005) did not test the network algorithm on a bigger dataset. Nevertheless, it can be shown that split networks computed by Network can be used to determine language contact due to shared innovations. However, the algorithm results in several disadvantages: the direction of transfer cannot be detected; the underlying evolutionary process can only be determined by linguists; and the amount of loanwords depends on the parameter setting. Additionally, an unrooted data-display network cannot make any assumptions on the ancestry of the languages. The network gives insights on the historical connections between languages independent of inheritance or transfer, i.e. explicit evolutionary processes like borrowing cannot be captured.

Evolutionary Networks

Evolutionary networks represent the evolution of the languages and horizontal transmissions according to evolutionary processes like hybridization or borrowing. Rooted networks can be derived from characters, clusters, or trees (Morrison, 2011). Similar to data-display networks, evolutionary networks aim to represent the complete evolution of the languages. The major difference is the visualization of the hierarchical structure indicating the classification of the languages according to ancestry. In an evolutionary network, the underlying clustering of the languages is kept, while reticulate edges indicating additional transfer are added.

The study of Willems et al. (2016) focuses on using hybridization networks to retrace the evolution of the Indo-European language family. In phylogenetics, a hybrid species is a mixture of two parental species combining the qualities of two organisms of different varieties through sexual reproduction. This means, one part of the hybrid genome comes from one parent and one part from another parent. An example would be a mule, which is a hybrid of a male donkey and a female horse. In linguistics, hybridization is similar to the emergence of mixed languages, i.e. pidgin or creole languages. A pidgin language develops between groups that have no language in common. It is a marginal language used to fulfill certain communication needs among those speakers without forcing one group to learn the language of the other (Campbell, 2013). A creole language arises out of a pidgin language, becoming the native language of the speakers in this newly formed community (Campbell, 2013).[20] Pidgin and creole languages emerge out of two mother languages during language contact instead out of one single ancestor language due to inheritance. They are special kinds of languages according to their emergence

[20] For a more detailed discussion and a list of pidgin and creole languages, see Campbell (2013, p. 309).

and evolution processes, which differ to the common processes of inheritance and horizontal transmission. Willems et al. (2016, p. 2) argued that "word borrowing can be viewed as the main development mechanism to the emergence of hybrid (i.e., mixed or contact) languages". In the emergence of a pidgin language, words are transferred from two languages to develop a new, simplified language to fulfill a specific communication purpose. The question arises whether the transfer of words to design a new language is allocated to borrowing or inheritance. In the case of inheritance, two mother languages pass on their linguistic material to the child, resulting in a new linguistic system compared to the inheritance from one single ancestor language. In the case of transfer, two languages transfer lexical material to create a new language compared to the modification of one language from another donor language. Both processes clearly differ from the common procedures known for existing languages. Schuchardt (1868) already mentioned that there is no totally unmixed language.[21] This means all languages are, in a sense, mixed languages, and the transfer of items to change them could be seen as a process of development, regardless of whether they are pidgin, creole, or existing languages. However, the adaptation and integration process of words to form a pidgin differs from the mechanisms known for loanword adaptation and integration into existing languages. An existing language contains inherited items from its mother language and borrows words from another donor language due to language contact. In contrast to the process of borrowing, where the mother and donor language are clearly identifiable, no distinction between mother and donor language is made in the case of pidgins. A pidgin language develops out of words transferred from two languages, where both are simultaneously mother and donor language. The borrowing procedure might seem to be similar; however, the requirements and outcomes are different and need to be visualized in separate networks. It is therefore questionable to adapt the parallel between hybridization and mixed languages proposed by Atkinson and Gray (2005) to all contact scenarios in linguistics and equate the two processes of horizontal transmission, as done by Willems et al. (2016). Although the reconstruction of hybridization networks to display horizontal transmission in the sense of borrowing is debatable in linguistic terms, Willems et al. (2016) proposed a solution where the reticulate events and their linguistic evolutionary processes can be explained and evaluated in terms of language contact. According to the linguistic evaluation in the study, the reticulations represent expected contact scenarios within the Indo-European languages, which can shed light on the evolutionary history in terms of horizontal connections between the languages.

The algorithm in the study of Willems et al. (2016) is based on lexical data and aims at detecting lexical hybrids and word borrowing events. The inputs of the algorithm

[21] Translated from the original: "Es gibt keine völlig ungemischte Sprache" (Schuchardt, 1868).

(Willems et al., 2014) are a distance matrix and three user-defined parameters, i.e. the algorithm is not parameter-free.[22] The result is an explicit hybridization network to model the evolution of the languages and to identify hybrid languages along with their corresponding parent languages. In addition, the algorithm computes a degree of hybridization for each reticulation, i.e. the percentage of transferred lexical material from their parents. The resulting network visualizes the language history, including the most prominent mixtures as reticulations. The network represents mixed languages in a sense that all languages show some degree of mixing due to language contact and borrowing. Languages with two parents are hybrids in a sense that they borrowed lexical items from both languages to a certain extent. The amount of transfer, i.e. the percentage of lexical items transferred from one language to the RL, is recorded at the corresponding reticulation. The direction of the transfer can be determined by the algorithm. However, since they are working with hybrids, no distinction between ancestral language and donor language(s) can be made, since hybrids are more similar to pidgin and creole languages than to existing languages. To account for this issue, Willems et al. (2016) proposed the usage of an underlying language tree, e.g. from Gray and Atkinson (2003), which displays the clustering of the languages due to common ancestry. The reticulations detected by the hybridization algorithm are added to the phylogenetic tree to construct an evolutionary network. The resulting network represents the common ancestry of the languages with respect to the underlying tree and two reticulations indicating horizontal transmission (hybridization) from two other languages.

Other evolutionary networks can be inferred from a set of two or more phylogenetic trees. Since the calculation of these networks is computationally hard, topological constrained classes of networks are introduced in phylogenetics to model evolutionary history. Among other rooted networks, a *galled network* is the most specific type of topological restricted network, next to the more general *level-k network*. In the literature, the term *level* refers to the maximum number of reticulations in a network, where k indicates the number of reticulations (Huson et al., 2010). A level-0 network is a phylogenetic tree, a level-1 network contains one reticulation, a level-5 network contains five reticulations, and so forth. Depending on the algorithm, the parameter k needs to be set manually or is selected by the algorithm according to the given data. The parameter restricts the reticulations which are present in the network. A level-k network cannot necessarily represent all sets of clusters compared to the more specific galled network, which is able to represent every representation of a given set of clusters. In a galled network, reticulation cycles can share edges, but

[22] Willems et al. (2016) give a short explanation on their selection of the parameters, but refer the reader to their original article (Willems et al., 2014).

they cannot share reticulation nodes (Morrison, 2011).[23] An algorithm to compute a galled network is implemented in *Dendroscope* (Huson and Scornavacca, 2012), a software to view and compute rooted phylogenetic networks out of trees.[24] The input for the algorithm to reconstruct a galled network displaying the evolutionary history of languages including horizontal transmission is a set of trees. Willems et al. (2016) constructed a galled network with Dendroscope using 200 concept trees, i.e. one for each concept in the basic vocabulary list.[25] The resulting galled network models hybridization events with two parents, i.e. the mother language cannot be distinguished from the donor language. The algorithm in Dendroscope computing the galled networks reclusters the objects to meet all topological restrictions. This can result in a different language clustering than proposed by the language tree. Without the usage of an underlying tree to classify the language relationships, both the hybridization and the galled network might result in a different language clustering. Delz (2014) showed that galled networks can also be reconstructed using one language tree and one concept tree to model borrowing. The resulting galled network is restricted to concept-internal borrowings, where the reticulations could indicate clear word borrowings between the languages. However, the disadvantages are still present and cannot be solved. A galled network models reticulations in case of hybrids, and not of borrowings between existing languages. The mother language cannot be distinguished from the donor language, the direction of borrowing cannot be determined, and the algorithm might change the language classifications due to inheritance. Consequently, galled networks are not an appropriate method to model language borrowing.

Neither hybridization nor galled networks can make a clear distinction between mother and donor language, which is important in a model of borrowing and loanword detection. The hybridization algorithm by Willems et al. (2016) can be merged with an attested underlying language tree, which is an unavoidable step to come closer to the common visualization of borrowing, where the mother language can clearly be distinguished from the donor language(s). This network displays the classifications of the languages according to the tree, and the most prominent horizontal edges according to the hybridization algorithm. The direction of the transfer and the reticulation degree are the major gains of the study. Although the network algorithm proposed by Willems et al. (2016) originates from a wrong parallel between phylogenetics and linguistics, it is a step in the right direction for a

[23] For a detailed description and a graphical comparison of the different networks, see Morrison (2011). Delz (2014) introduces the different topologically restricted networks in terms of linguistic transfer scenarios.

[24] Dendroscope (Huson and Scornavacca, 2012) also contains an algorithm to compute level-k networks and other types of topologically restricted networks.

[25] Please see Willems et al. (2016) for a detailed explanation on the computation of concept trees used in their study.

better understanding of the evolution of natural languages, including both common ancestry and horizontal transmission.

In linguistics, the *minimal lateral networks* (MLN), originally introduced in phylogenetics by Dagan and Martin (2007), were applied to the task of automated borrowing detection (Nelson-Sathi et al., 2011). The underlying method to compute a MLN is one of the best-explored network methods in linguistics, and was applied to several datasets, including Indo-European (Nelson-Sathi et al., 2011; List et al., 2014a), Chinese dialects (List et al., 2014b; List, 2015), and Austronesian (Jäger and List, 2018). Although MLNs are mainly used to model language contact, the method can give insight into the underlying lexical flow causing the contact scenario. The linguistic model of *phylogenetic lexical flow inference* (PLFI), developed by Dellert (2019a), uses information-theoretic causal inference to discover language contact. The main aim of the algorithm is to reconstruct contact scenarios and generate an evolutionary network, including reticulate edges, to illustrate the lexical transfer. In addition, the detected contact situations in a PLFI analysis contain additional information on specific borrowing events and the lexical items involved in the transfer. Both the MLN and the PLFI algorithm can be used to extract more information on borrowings and loanwords. Hitherto, less attention was paid to the detection of loanwords using these algorithms. In chapter 5, the algorithms, their adaptation to the task of loanword detection, and their (dis)advantages are introduced in more detail.

3.3 Horizontal (Word) Transfer

HGT networks capture evolutionary events which arose due to a specific kind of transfer, i.e. the horizontal transfer of genes. During the HGT process, genes are transferred from one organism to another without sexual reproduction (Morrison, 2011). The transfer of genes causes changes within the genetic structure of the organism, which leads to a modification of a single lineage. Each gene transfer is unique, and depends on the contact situation where genetic material is added from a donor (Morrison, 2011). The network should therefore be able to represent the direction of transfer, the gene under consideration, and the donor and recipient. To ensure for these results, the underlying algorithms are adjusted according to the theory and methodology of horizontal gene transfer.

The process of borrowing is comparable with the process of HGT. Two languages are in contact with each other and exchange linguistic material, which results in a modification of the recipient language. Different types of linguistic material can

be transferred, which is restricted by the intensity of the contact situation, i.e. the borrowability of linguistic material can be measured on the degree of contact. Since the lexicon is one of the language's less stable parts, lexical items can be transferred easiest between both related and unrelated languages. Lexical items can be seen as the genes of a language, which both form the basis of the language as a whole and are objects of transfer between two languages in contact situations. During the process of borrowing, the speakers involved in the contact situation decide on the lexical items which are transferred and included in their language. In accordance with the process of HGT, a network of horizontal word transfer (HWT) should represent the direction of borrowing, the loanword, and the donor and recipient language. This can only be captured within the framework of horizontal gene transfer.
For a complete understanding of the underlying methods and algorithms used later in this thesis, the theory and concept of horizontal gene transfer is introduced. The parallel to linguistics according to the concept of HGT is explained to specify the idea in linguistic terms. Lastly, linguistic studies using the idea and methods introduced in phylogenetics for HGT are discussed.

3.3.1 HGT: Fundamental Concept and Methodology

Since HGT occurs when a gene is transferred by other means than sexual reproduction, different processes can be involved including *transformation*, *conjunction*, and *transduction* (Morrison, 2011). In case of transformation, the genetic material is incorporated from the surroundings of the cell. The process of transduction involves e.g. a virus, where the virus functions as a messenger carrying the genetic material from one cell to another. Conjunction is the only process where the cells are in direct contact with each other exchanging material. All of these processes modify one gene in the cell and change the evolutionary history of this gene due to horizontal transmission. It is therefore obvious that the phylogenetic tree of a gene or genome fragment involved in HGT will differ from the classifications represented in the species tree.
The characteristics of horizontal gene transfer are the following (Morrison, 2011):

(i) HGT does not happen during sexual reproduction.
(ii) HGT can only occur between contemporaneous taxa.
(iii) HGT modifies a single lineage by adding new genetic material.
(iv) HGT can affect both only one gene or various genes.
(v) HGT can affect closely and distantly related lineages.

These characteristics distinguish HGT from other evolutionary processes, like e.g. hybridization, deep coalescence, and duplication–loss scenarios, which happen due to sexual reproduction and can be explained by regularities (Morrison, 2011). The modification of a single lineage and the impact on one or more genes due to HGT affect the evolutionary history of the gene, but not the development of the species due to inheritance. The recognition of possible discordance between the evolutionary history of a gene and its species has led to proposals for treating genes as single characters to model their explicit evolutionary history using phylogenetic trees (Page and Charleston, 1997). The unique history of a gene is therefore represented in a gene tree, whereas the species tree represents the overall evolutionary history of the species.[26] The underlying idea to detect HGT events is the reconciliation of a gene tree with its species tree to explain the history of a gene by fitting its lineage into the branching pattern of the species tree. The incongruence between the two trees can be captured by introducing a specific number of evolutionary events. The reconciliation of trees takes the other HGT characteristics into account. The effect on closely and distantly related lineages can be captured, the merging of the trees can account for differences between contemporaneous taxa, and the mismatches can be computed and displayed as reticulations.

The idea of tree reconciliation dates back to Goodman et al. (1979), who aim at detecting evolutionary processes by fitting gene trees to their corresponding species trees. Page (1994a) formalized the idea of Goodman et al. (1979) to map a gene tree into its species phylogeny by describing an algorithm and developing methods to measure the degree of fit between the trees. One of the main gains of the study is the visualization of a so called map of trees to display the reticulations.[27] Further algorithms have been developed to detect transfer events by counting the minimum number of events needed to fit the gene tree into the species tree. Most of the methods only need tree topologies as inputs.[28]

For the detection of mismatches between two trees, the methods need a species and a gene tree as input. In comparison, hybridization algorithms take multiple gene trees as input to merge them into a species network with additional reticulate edges. However, since horizontal transmission affects only one gene, only one of many gene trees will reflect this pattern. The algorithms used to detect hybridization are consensus methods, and are therefore likely to ignore a pattern in a single tree. The

[26] Page and Charleston (1997) mentioned that the usage of gene trees shifts the attention from the relationships between characters and trees to the level of relationships between trees.

[27] In further studies, the idea of reconciling trees is used for different purposes. Page (1994b) used a reconciliation method to detect host switching events caused by the process of transduction. Mirkin et al. (1995) and Maddison (1997) developed alternative methods for combining gene trees to reconstruct a single species tree.

[28] For an overview of available software to detect hybridization and horizontal gene transfer, see Morrison (2011, p. 159).

advantage of mapping one gene tree into its species tree is the direct detection of transfer in the history of this single gene. This results in the explicit identification of the gene, the transfer, and the respective donor and recipient. Depending on the algorithm, the direction of transfer can be added to the model. Figure 3.5 shows an illustration of a reticulate edge due to horizontal transmission. The first gene tree

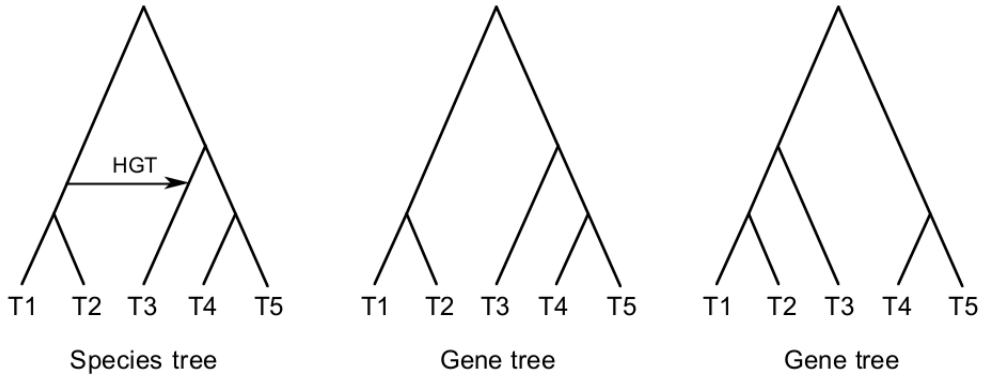


Fig. 3.5.: A representation of a species tree (left) with five taxa (T1–T5), with two taxa involved in an HGT event, plus two gene trees. The first gene tree (middle) reflects the history of the species tree, whereas the other gene tree (right) differs from the history of the species tree due to horizontal transmission. The graphic is from Morrison (2011, p. 112).

(middle panel) is in accordance with the species tree, i.e. the evolutionary history of the gene and species are the same. This results in a perfect match of the two trees, where no reticulations are needed to explain the difference between them. The evolution of the second gene (right panel) differs from the history of the species. To map the gene tree into the species tree, the branch of taxa T3 in the gene needs to be moved to another position to correspond to the branching pattern of the species tree. This is indicated by adding a reticulate edge. The evolutionary history of the species is kept in the resulting network, and the horizontal transfer by which the gene is affected is displayed by a reticulation.

However, the reconciliation of evolutionary histories is mathematically NP-hard, which makes it problematic to develop exact algorithms for reconciliation of trees. This leads to the problem that most algorithms can only detect potential transfer events, which could lead to the addition of arbitrary events. Additionally, there are two ways of interpreting the output of a HGT network (Morrison, 2011). First, the output could be explicit, i.e. the network is an interpretable diagram. Second, the output could be implicit, i.e. insight is needed to interpret the diagram. In most studies, the resulting evolutionary networks are treated as explicit networks, where each reticulation is treated as an evolutionary event due to horizontal transmission. Morrison (2011) argues that the networks should be treated as implicit. This is

that mathematically, every hybridization or HGT event creates a reticulation cycle in the network, i.e. a cycle with three or fewer outgoing arcs. The interpretation of the cycles have a great impact on the biological explanations of the networks. This means, in some cases it is difficult to distinguish the various causes of evolutionary events which could cause the reticulations. In biology, hybridization and horizontal gene transfer are closely related to each other. Hybridization is the exchange of genes due to sexual reproduction, mostly present in animal or plant breeding. Horizontal gene transfer is the exchange of genetic material between unicellular and/or multicellular organisms like bacteria. An implicit interpretation of a network can therefore only be done by experts in the field, and not by the algorithm itself.
Biologically, both the mapping of a gene into a species tree and the summarizing of gene trees into a species network can be applied to detect reticulate events. Morrison (2011) mentioned that it is a historical accident that gene-tree reconciliation is so strongly associated with the detection of horizontal gene transfer. However, from a linguistic point of view, this association is important, since the distinction between the evolutionary processes of hybridization and horizontal transfer plays a crucial role in language evolution.

3.3.2 The Concept of HGT in Linguistics

During language contact, linguistic material is transferred from one language to another. The transfer can be caused not only by a large number of different underlying processes, but also by several social and linguistic factors. The composition of the processes and the factors included is unique, and can differ depending on the contact situation. Further, the contact persists during a certain period until it breaks off. The challenge lies in both locating the contact phase and in detecting the transfers, including the surrounding factors and the underlying processes. It is not surprising that the identification of the complete contact situation is difficult, if not impossible. The focus therefore lies on the detection of loanwords, which remain in the RL as result of the borrowing process. Although loanwords are adapted and integrated into the recipient language, they are the most detectable items to identify a borrowing scenario and to trace back their donor language in order to specify the contact situation.
The process of borrowing shows similar characteristics to the process of HGT.

(i) Borrowing does not happen due to inheritance.
(ii) Borrowing can only occur between contemporaneous languages.
(iii) Borrowing modifies a single language by adding new linguistic material.
(iv) Borrowing can affect only one word or various words.

(v) Borrowing can affect closely and distantly related languages.

Those characteristics are distinctive for the processes of borrowing, and clearly show the distinction to the process of hybridization, where two languages are mixed to create a new one. The process of borrowing is not due to inheritance. In linguistic terms, this means that the speaker of a recipient language borrows words to fulfill particular needs in order to enhance their language for a better communication. The borrowings are specified by the speech community and are primarily caused by social constraints, like need or prestige, and by various linguistic constraints. This implies that borrowing can only occur between contemporaneous languages, since languages can only be in active contact with each other during the same time period. There are neither restrictions on the intensity and degree of contact, nor on the transfer process itself, which includes borrowing between both closely and distantly related languages as long as they are in contact with another. During language contact, the recipient language borrows from the donor language, i.e. only the RL is modified by the transfer of lexical items. Each modification is specific to the borrowing process, the donor language, and the transferred items. The amount of words transferred from the SL to the RL is determined by the speakers involved in the transfer, while the whole speech community is responsible for the integration and maintenance of the new lexical items. It is therefore not surprising that borrowing affects the evolutionary history of the word involved in the transfer. The modification of the word's history, however, does not have a great impact on the evolutionary classification of the language itself within the language family.

The similarities between HGT and borrowing are obvious, which leads to the adaptation of the corresponding methods for tree reconciliation to linguistics. Since the evolutionary history of a word is represented in a concept tree, and one of the languages in a language tree, the trees are equivalent to gene and species trees in phylogenetics and serve as input for the tree reconciliation methods. The mapping of a concept tree into its language tree provides insights into the evolution of the lexical items to detect contact and transfer situations. The discordance between the trees is measured by the number of evolutionary events, which need to be introduced to fit the concept tree into the language tree. This analysis takes all borrowing characteristics into account, i.e. borrowing between closely and distantly related languages can be captured, as well as the modification of a RL through another language.

A concept tree captures transferred lexical items within one concept which have an impact on the evolutionary history and modify the recipient languages due to borrowing. Since the pattern is likely to occur only in one single concept tree, the direct mapping of the concept tree onto its language tree gives insights about the

discordance between the trees with respect to the concept under consideration. This results in a detection of evolutionary events for a specific concept, including the transfer, the respective donor and recipient language, and the direction of transfer. The obvious discordance between the concept tree in figure 3.3 and the language

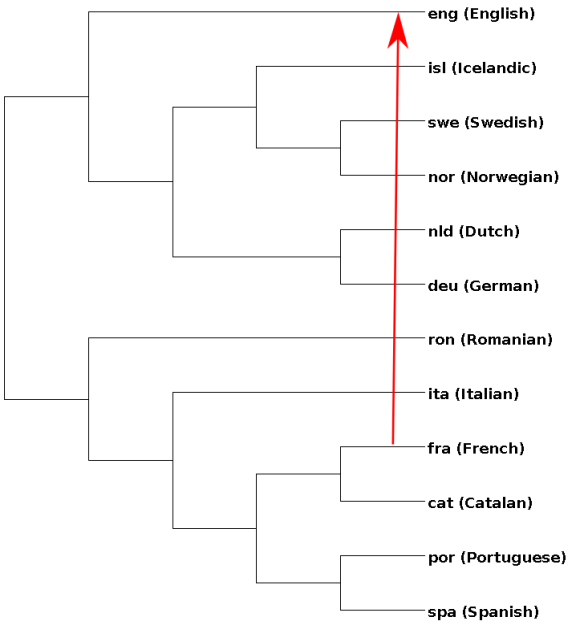


Fig. 3.6.: The sketch of a visualization of a network to model horizontal word transfer.

tree in figure 3.2 is the position of English. The concept tree displays the evolutionary history of the word 'mountain', where English is grouped within the Romance languages. This reflects the attested borrowing of the word *mountain* from Old French (Merriam-Webster Dictionary, 2020). Figure 3.6 displays a sketch of a HWT network visualizing the borrowing between French and English. The language tree is the underlying model of the network, keeping the language clusters according to the classification of the computational comparative method. The detected evolutionary event is added to the language tree in form of a reticulate edge between the two languages. The arrow indicates the direction of transfer between the SL French and the RL English. The tree reconciliation methods from phylogenetics aim at detecting these kinds of evolutionary events, where the loanword can clearly be identified in addition to the transfer, including the donor and recipient language.

The adaptation of phylogenetic methods for tree reconciliation does not circumvent the problem of being mathematically NP-hard. In linguistics, the application of the methods leads to the same problem, namely the detection of potential HWT events by most algorithms. The resulting networks could therefore contain arbitrary events which might not be due to borrowing. In addition, the interpretation of the

output, mentioned by Morrison (2011), does also hold for linguistic data. Networks of horizontal word transfer can be either interpreted as explicit or implicit. The explicit interpretation is an interpretable diagram, where each reticulation is treated as transfer event. The implicit interpretation requires linguists and their expert knowledge to distinguish arbitrary from real evolutionary events by identifying true borrowings within the network. This kind of interpretation cannot be done by the algorithm. In phylogenetics, the two processes of hybridization and HGT are closely related, where a clear distinction of the events is needed for a complete biological analysis. In linguistics, hybridization should be modeled in a different network than borrowing, since the two evolutionary processes need to be analyzed separately. An explicit analysis gains in significance, since tree reconciliation methods are used exclusively for the identification of borrowings. However, the detection of potential HWT events, along with the appearance of arbitrary events, ensures for an additional implicit analysis to underline the results.

In linguistics, the difficulty for a proficient interpretation of HWT events lies within the complexity of the adaptation and integration processes of loanwords. Those processes are individualized and differ according to the languages included in the contact scenario. The results of an optimal computational analysis should reflect the state of a loanword within a specific time period. For the concept and language trees computed using the basic vocabulary list of NELex, it can be assumed that all loanwords are fully integrated in the languages. The borrowed word still needs to show a different pattern than the other words to be identified as loanword. However, a complete integration of a word in a language makes it difficult to analyze a pattern indicating the status of a word as loanword. On the one hand, Kessler (2001) already mentioned this uncertainty of analyzing a loanword at a specific point in time, since the older the loanword, the more it is integrated in the RL. On the other hand, Haugen (1950) introduced different types of loanwords and their integration processes, which could lead to different patterns in loanwords compared to inherited words. A tree reconciliation algorithm cannot analyze such finely graded patterns of loanwords and answer the question of identifiability. The results of the algorithms depend highly on the data and the reconstructed input trees. All methods leave some room for interpretation by experts to clarify the detected contact scenarios and the true borrowings within the network.

The identification of mismatches between two trees in order to detect evolutionary events due to horizontal transmission opens a new research field. The phylogenetic algorithms give insights in the evolutionary history of lexical items and their relation to the language classifications proposed by the computational comparative method. The methods provide a straightforward way to discover language relationships, contact scenarios, and borrowings. Boc et al. (2010a) used a HGT algorithm

(Makarenkov et al., 2006) to represent the classification of the Indo-European languages within a phylogenetic network. The reconstructed distance-based concept trees are mapped to the language tree proposed by Gray and Atkinson (2003), resulting in different tree reconciliation scenarios. All results obtained from the borrowing scenarios are combined to compute a borrowing statistic.[29] The language tree from Gray and Atkinson (2003) is extended by the ten most prominent transfer events to create a network. The HGT network illustrates the relationships of the Indo-European languages, including the ten most prominent contact scenarios indicated by reticulate edges. The focus of applying tree reconciliation methods to detect language contact lies in the identification of language relations within an evolutionary network. Borrowing scenarios and the detected loanwords were not analyzed in the study. In a preliminary study, Delz (2013) used the algorithm by Boc and Makarenkov (2003) to illustrate the detection of borrowings and loanwords within the HGT framework. The results are displayed in an evolutionary network, where the language tree provides the underlying structure and the transfer events detected by the algorithm are added as reticulate edges. Delz (2013) evaluated the results of the detected evolutionary events manually and contextualized the linguistic analysis of the contact scenarios. In a follow-up study, Delz (2014) developed a tree comparison method based on the idea of jackknifing to detect loanwords using mismatches between a language and a concept tree. The tree comparison algorithm is the first linguistic method using a language and concept tree as input to discover borrowings and loanwords, which is why it serves as a comparative algorithm to the HGT algorithms in this thesis.[30] The generated evolutionary networks including borrowings and loanwords are evaluated manually, since gold standard data for the database under consideration was not available for a statistical evaluation. The main aim of the study was to elucidate the role of donor and recipient languages within the borrowing process, along with the complexity factor of the determination of both RL and SL in the contact scenario and the network. To my knowledge, no further cross-linguistic studies on borrowing including a detailed loanword analysis using tree reconciliation algorithms have been attempted. The main aim of this thesis is to clarify the purpose of tree reconciliation methods in linguistics. The following analyses should give insights to which degree the direct transfer of phylogenetic methods into the field of linguistics is fruitful and can be used to discover borrowings and the corresponding loanwords. The detection and identification of loanwords is a first step into the direction of a deeper understanding of language contact, contact scenarios, and the possible types of loanwords present in the data. The adaptation

[29] For a detailed representation of the borrowing statistic, please see the tables and explanations in Boc et al. (2010a).

[30] The idea of jackknifing and the algorithm are explained in detail in chapter 5.

of phylogenetic methods is not only worthwhile to shed light on detailed horizontal transmissions, but serves as basis for further, more detailed analyses in the field of loanword detection.

4 Concept Trees: From Wordlists to Word Trees

Trees are the major foundation of both network and tree reconciliation methods. In order to identify horizontal word transfer within the framework of tree reconciliation, both concept and language trees are needed for the analysis. The computational comparative method results in a language tree which displays the language classifications according to their shared innovations. The methods for automatic language tree reconstruction are well-tested and evaluated by several linguistic studies, which present a high agreement between automatically reconstructed trees and expert classifications (see e.g. Brown et al. (2008), Holman et al. (2008), Jäger (2013a), and Jäger (2013b)). In contrast, research in CHL paid less attention to the reconstruction of concept trees to illustrate the evolution of individual words within the tree model. Up to this point, only a few studies have made use of concept trees in linguistics to address the research question of detecting horizontal transmission (Delz, 2013; Delz, 2014; Willems et al., 2016). The concept trees were reconstructed to serve the purpose of reconstructing evolutionary networks and discovering transfer events, i.e. they are a secondary effect of the studies. The concept trees in these studies are reconstructed using simple sequence alignments and distance-based methods. No further research on the application of other phylogenetic methods for matrix computation, tree reconstruction methods, or evaluation techniques in terms of concept tree generation was made. One challenge is to adapt suitable methods from phylogenetics into linguistics. Besides this, the methods need to be adjusted to fulfill linguistic requirements important for concept tree reconstructions. However, in terms of gene tree generation, phylogenetic methods are not as well established as it might seem. Gene trees are mainly reconstructed for the computation of networks or consensus trees in order to explain the evolution of the species (Liu and Pearl, 2007; Nakhleh et al., 2009; Rasmussen and Kellis, 2007; Doolittle et al., 1996). In addition, some studies are using small datasets to detect horizontal transfer using gene trees (Hallett and Lagergren, 2001; Boc and Makarenkov, 2003; Nakhleh et al., 2005b; Makarenkov et al., 2006). Nevertheless, few research is done on matrix computations for gene data, and very little in terms of gene tree evaluation. These issues will be addressed in this chapter with regard to linguistic data and concept tree reconstruction.

In terms of borrowing detection, the measure of surface similarities has been proven useful and can serve as a proxy for the detection of borrowings (Van Der Ark et al., 2007). Loanwords show a high degree of surface similarity with their corresponding counterparts in the SL from which they were copied (Haugen, 1950; List, 2019). The comparison of the words can give insights into the relationships among them and possible transfer scenarios between languages. It is therefore necessary to have a closer look at the phylogenetic methods for sequence comparison and matrix computation to establish suitable methods which can be applied to linguistic data. Some of the methods can be applied with small adjustments, whereas other methods need to be customized in order to handle linguistic data in an appropriate way. In addition, tree reconstruction methods are revised and adapted for the computation of concept trees. This is the initial step in order to gain reasonable results in terms of borrowing and loanword detection using tree reconciliation methods. It is therefore essential to address the issue of concept tree reconstruction to elaborate suitable methods for linguistic data.

The underlying procedure for concept tree reconstruction was introduced in section 3.1.4, which will be extended in terms of new ideas and innovations to compute matrices and generate trees. The tree reconstruction methods are divided into distance-based and character-based methods respectively, where the best attested methods for tree reconstruction in linguistics are adapted to model concept trees. In order to achieve usable results, the methods for matrix computation are adapted as well to fulfill the requirements for concept tree reconstruction.

First, the distance-based methods are introduced, including different matrix generation methods and sampling techniques to assess tree replicates and the stability of inner clades for evaluation purposes. Second, the computation of data matrices for character-based methods are presented, including the sampling techniques of the algorithms to extract tree replicates. For all distance-based and character-based methods, the ideal method is determined. The concept trees computed by the most optimal methods serve as input for the loanword identification task.

4.1 Distance-Based Methods

Distance-based methods were one of the first tree reconstruction methods adapted to CHL from phylogenetics.[1] The major advantages are their high efficiency and their application to a wide range of different data types (Huson et al., 2010; Jäger and List, 2016). The underlying distance matrix can be computed using different

[1] For a digression on the history and philosophy of phylogenetic methods, see Felsenstein (2004).

kinds of measurements, where sequence alignment is the most widely used. The input for the alignment are lists of lexical items, which can vary in amount and classification of the words, and depend on the dataset under consideration. Since, the comparison of words is the standard technique to classify languages according to the (computational) comparative method, linguists laid a stronger focus on the collection of word lists for various languages to optimize their classifications. This results in a wide range of databases providing different kinds of word lists covering various language families. The availability of databases covering a great amount of language families in the world is the reason why distance-based methods are quite popular to reconstruct linguistic trees (Jäger, 2013a).

Word lists are the starting basis for the computation of distance matrices. Based on the datatype and the transcription of the lexical items, a distance measure can be developed to reconstruct a linguistic tree. This means various distance measures can be applied to the same data, which results in a greater range of experiments. Since the reconstruction algorithms are computationally efficient, they can also be applied to large distance matrices in a reasonable time (Huson et al., 2010). This is a great advantage, since the amount of data only has influence on the alignment and matrix computation, and not on the reconstruction of linguistic trees. Therefore, language trees for one or more language families can be reconstructed to shed light on the classifications of the languages and their relationships.

The underlying concept and procedure for concept tree reconstruction is introduced in chapter 3.1.4. The input data in form of word lists is taken from the NorthEuraLex database, introduced in section 2.2.3. The unaligned lexical items are compared using sequence alignment methods to compute distances represented in matrices. Within the task of concept tree reconstruction, one challenge is the adjustment of alignment methods for the matrix generation. There are different ways to compute them, using phonetic representations or cognate classes of lexical items. Furthermore, geographical language distances can be added as additional parameters to the estimation of the distances between word pairs. The general idea is to integrate additional components, e.g. cognate classes and geographical distances, into the alignment analysis to improve the distance measurement. Both traditional sequence alignment techniques and newly developed methods are described in the next subsections. The matrices serve as input for the tree reconstruction methods. In chapter 3.1.2, three different reconstruction algorithms are introduced and compared. In phylogenetics, the FastME algorithm (Desper and Gascuel, 2002; Desper and Gascuel, 2004) is the computationally most efficient algorithm, resulting in more accurate trees than the other two algorithms. For the reconstruction of linguistic trees, Jäger (2013a) compared the distance-based algorithms implemented in the *FastME* software package (Desper and Gascuel, 2002). FastME provides

BME-based reconstruction algorithms, including post-processing techniques for tree improvement. In terms of automated language classification, the BIONJ algorithm, introduced by Gascuel (1997), performs best.[2] Additional post-processing methods improve the results significantly. Following the results of Jäger (2013a), the FastME algorithm, implemented in the software package *ape* (Paradis et al., 2004) with post-processing techniques is used to reconstruct concept trees in the following analyses.

4.1.1 State-of-the-Art: Pairwise Sequence Alignment and Distance Matrix Computation

Sequence comparison in linguistics aims at identifying shared patterns between the phonological representations of two lexical items. Alignment methods from phylogenetics represent the string comparison in a matrix and compute either a similarity or distance score between two lexical items. The methods are adapted into CHL for the computation of similarities or distances between each pair of lexical items in the dataset.

In linguistics, the *Levenshtein distance* is the standard approach for pairwise sequence alignment. The method computes a distance score between two sequences, where the distance is defined as the minimal number of edit operations needed to transform one sequence into the other one. Holman et al. (2008) proposed a normalized version of the Levenshtein distance to account for varying word length. The Levenshtein distance is normalized by dividing the distance score by the length of the longest word.

The normalized Levenshtein distance can only distinguish between identical and non-identical sounds (Jäger, 2013b). In order to achieve a better approximation of the etymologically correct alignment, similarity values of segment pairs can be taken into account to estimate the probability of sound correspondences. The *weighted Needleman-Wunsch* (NW) algorithm (Needleman and Wunsch, 1970) uses a given substitution matrix and gap penalties to efficiently generate the optimal alignment and its similarity score.[3] The substitution matrix reflects the similarity scores between sounds, which are estimated in CHL using *Pointwise Mutual Information* (PMI). Jäger (2013b) developed a PMI-based method for the computation of string similarities using the ASJP database. Within the PMI-based framework, the NW algorithm efficiently determines the optimal alignment given the PMI values and

[2] See Jäger (2013a) for a detailed linguistic evaluation of the algorithms.

[3] The Needleman-Wunsch algorithm is a generalization of the Levenshtein distance. For a comparison of the two measurements, see (Jäger, 2013b).

gap penalties. The PMI scores estimate the probability of sound correspondences between all segment pairs across the 41 symbol encoding of the ASJP database.[4] The computation of the PMI scores is implemented in the *OnlinePMI* program by Rama et al. (2017), which was used to calculate the PMI scores for NELex using the ASJP encoding system. When aligning lexical items, there are often gaps indicating that a sound has no corresponding counterpart in the other word, as shown in the repeated alignment example 4.1 below.

(4.1) Pairwise alignment of the words for the concept 'mountain'

a. French: m o - t a 5 -
Spanish: m o n t a 5 a

b. English: m a u n t - i n
French: m - o - t a - 5

c. English: m a u n t i - n -
Spanish: m o - n t - a 5 a

d. English: - m a - - u n t i n
German: b - E r k - - - - -

e. German: b E r k
Dutch: b E r x

The NW algorithm uses one parameter for opening a gap, i.e. insertion of a new gap in the alignment, and one for extending the gap. The gap penalties optimize the alignment and the resulting alignment score. The proposed gap penalties by Jäger (2013b) are used futheron in the analyses for two reasons. First, appropriate values for both parameters can only be found via optimization, carried out in a precise way by Jäger (2013b) for the ASJP encoding system. Second, since Rama et al. (2017) used the proposed gap penalties within their OnlinePMI algorithm for estimating the PMI scores, the same parameters are applied to other analyses in order to stay consistent in parameter selection. Using a substitution matrix of PMI scores and gap penalties, weighted NW improves the accuracy of the distance measure compared to the Levenshtein method (Jäger, 2013b). Therefore, the PMI-weighted NW algorithm is the method of choice for the computation of word pair alignments. However, the method results in a PMI-based similarity score between each pair of lexical items. Since the distance-based tree reconstruction algorithms require a distance matrix as input, the similarity scores need to be converted into distances.

[4] See Jäger (2013b) for a detailed description of the parameter estimation.

Distance d_{dow} The *ALINE* system introduced by Downey et al. (2008) provides a formula to transform similarities into distances.[5] The similarity is converted into a distance score by normalizing it with the similarity score of the word self-comparison:

$$d_{dow}(w_1, w_2) = 1 - \frac{2 \times s(w_1, w_2)}{s(w_1, w_1) + s(w_2, w_2)} \tag{4.1}$$

where $s(w_1, w_2)$ is the similarity score resulting from the NW alignment between the two words w_1 and w_2. The word self-comparison of the two words is presented by $s(w_1, w_1)$ and $s(w_2, w_2)$ respectively. The aggregation method can be extracted and used to compute a distance score from the NW alignments within the PMI-based method. The distance score directly represents how distinct or similar two lexical items are to another. Downey et al. (2008) note that this approach is a useful measurement of phonetic distances and can therefore be used to generate distance matrices for automatic tree reconstructions.

Sigmoid Transformation Another approach to transform a distance score using NW alignment was proposed by Rama et al. (2017). Within their OnlinePMI system, they used the *sigmoid transformation* to convert a similarity score into a distance score:

$$d_{sig}(w_1, w_2) = 1 - \left(\frac{1}{1 + exp(-s(w_1, w_2))} \right) \tag{4.2}$$

where $s(w_1, w_2)$ is the similarity score between two words, w_1 and w_2, computed using the Needleman-Wunsch algorithm. The distance scores resulting from the sigmoid function lie in the range of $[0, 1]$. Within the multilingual cognate clustering task of Rama et al. (2017), word pair distance in the range of $[0, 1]$ can be used to estimate a cutoff point in order to determine the cognacy of two words. The classification of cognates according to their PMI score serves as input for the clustering algorithm, which leads to more accurate results in terms of automated cognate clustering (Rama et al., 2017).[6] To my knowledge, the sigmoid transformation and the resulting distances are until now not used for distance matrix generation and

[5] ALINE was developed for linguistic analyses, where each sound is represented as a vector of manually collected phonetic features. The distance between two sounds is determined by their difference in feature values. A comparison between ALINE and the PMI-based method can be found in Jäger (2013b).

[6] For more details on the automatic cognate clustering and the comparison between other methods, see Rama et al. (2017).

concept tree reconstruction.
Both distance transformations, d_{dow} and d_{sig}, are used to generate distances matrices for the tree reconstructions. In the distance matrix, the entries along the diagonal represent the distance of same words, which are 0, while the off-diagonal entries represent the distance between words from different languages. A sample distance matrix for the concept 'mountain' was given in table 3.5.

Handling Synonyms

For some languages and concepts, the NorthEuraLex database contains synonyms, i.e. the languages have more than one entry for a specific concept in the word list. An excerpt of NELex for the concept 'mountain' was given in table 3.4, which is repeated in table 4.1 below. Compared to the other languages, Swedish and Norwegian contain two words for the concept 'mountain'.

	Meaning 'mountain'	
	word	cc
English	`mauntin`	`m`
Icelandic	`fEtl`	`f`
Swedish	`bErg, fyEl`	`b,f`
Norwegian	`bErg, fyEl`	`b,f`
Dutch	`bErx`	`b`
German	`bErk`	`b`
Romanian	`munte`	`m`
Italian	`monta5a`	`m`
French	`mo,mota5`	`m`
Catalan	`munta53`	`m`
Portuguese	`mot3, mota5a`	`m`
Spanish	`monta5a`	`m`
...		

Tab. 4.1.: An excerpt of the data from the NorthEuraLex Database containing the phonological representations of the concept 'mountain' in the ASJP alphabet and manually assigned cognate classes.

The most prominent way of dealing with synonyms is to choose one word at random for the analysis. In this case, one word is chosen as representative for Swedish and one for Norwegian, whereas the other word(s) are excluded from the analysis. As a result, each language is represented by one word, i.e. each language is present once in the set of leaves of the concept tree. To achieve a better estimation of the language classification, the synonyms should be included in the analysis. The number of representatives of a concept can have an effect on the similarity between two languages, and therefore also on the language clustering in the tree. In addition, taking all words into account is in accordance with the linguistic perspective of the comparative method.

One strategy is to split the set of words for the languages into single entries in the word list, i.e. the number of synonyms determines the number of entries for the language. The languages containing synonyms would be present more than once in the distance matrix, and therefore also in the reconstructed concept tree. In the above example, Swedish and Norwegian would appear twice in the dataset: $Swedish_{bErg}$, $Swedish_{fyEl}$, $Norwegian_{bErg}$, $Norwegian_{fyEl}$. Each entry is represented in the distance matrix and in the language clustering of the tree. However, in order to reconcile a language tree and a concept tree for the detection of horizontal transmission, both trees need to have the same set of leaves. If the concept tree contains one language entry for each lexical item present in the corresponding language, the synonyms would cause a tree with a greater set of languages on the leaves. This means the number of languages can differ between several concept trees, since the number of leaves depends on the contained synonyms within each concept. The language tree, however, is a representative of the language classifications according to the comparative method, i.e. each language is present once in the tree. Since each concept tree should be reconciled with the same language tree, all concept trees need to have the same set of leaves as the language tree. In order to integrate synonyms in the analysis, it is essential to find a solution which can be implemented in the distance-based framework.

In the following approach, synonyms are integrated into the distance computation in such a way that both requirements, the consideration of synonyms and the single occurrence of a language in the concept tree, can be realized. The initial idea was presented by Willems et al. (2016) to account for multiple meanings while computing distances between cognates. The method is adapted to the task of distance matrix computation via pairwise alignment for concept tree reconstruction. Following the general idea of the method, an overall distance of a language pair containing synonyms can be computed. For two languages, all possible combinations of word pairings are aligned, and a distance is computed for each alignment. From the resulting set of distances, the minimal score is chosen as representative of the languages' distance for this concept. The method is formalized in the following way:

$$s_n^{l_i} = \{w_m^l | n = m \& l_i = l \& l \in L\} \tag{4.3}$$

$$d_m(l_i, l_j) = min\{d(x) | x \in s_n^{l_i} \times s_n^{l_j} \& l_i \neq l_j\} \tag{4.4}$$

Let L be the set of languages, W_m be the set of words for meaning m, and $w_m^l \in W_m$

a word in W_m belonging to language $l \in L$. Since there can be synonyms, a set of words for each language is constructed where $s_n^{l_i}$ is the set of words including all lexical items for meaning m of language l. The distance $d(x)$ between a pair of words x is computed using the PMI-based NW algorithm. The similarities are converted into distances using the respective methods d_{dow} (equation 4.1) and d_{sig} (equation 4.2). The minimum value of the set is taken as the representative distance between the two languages. If two languages under comparison contain no synonyms, i.e. one lexical item for the concept, the set for each language $s_n^{l_i}$ is a singleton. The distance between the languages is therefore the converted PMI-based similarity score from the NW algorithm. The result of the method is the distance $d_m(l_i, l_j)$ between two languages l_i and l_j for meaning m. For each concept, a distance matrix is generated, out of which a concept tree is reconstructed where each language is present only once.
Having a look at the sample data in table 4.1, three possible scenarios of language comparisons can be extracted:

1. A comparison of two languages where both contain synonyms: Swedish and Norwegian.
2. A comparison where only one language contains synonyms: Swedish and German.
3. A comparison where no language contains synonyms: German and Dutch.

In the first example, the set of languages is $L = swe, nor$. The set of words for each language is $s_n^{swe} = bErg, fyEl$ and $s_n^{nor} = bErg, fyEl$, respectively. The distance is computed by taking the minimum of all distances in $d_m(swe, nor) = min(d(bErg - fyEl), d(bErg - bErg), d(fyEl - fyEl))$, i.e. the minimum distance out of three is chosen as representative of the comparison. In the second example of $L = swe, nor$, only Swedish contains synonyms. The set of words are $s_n^{swe} = bErg, fyEl$ and $s_n^{deu} = bErk$. The distances are computed using $d_m(swe, deu) = min(d(bErg - bErk), d(fyEl - bErk))$ and choosing the minimum score out of two possible ones. In the last example, each language in $L = deu, ndl$ has one lexical item for the concept, resulting in the set of words $s_n^{deu} = bErk$ and $s_n^{nld} = bErx$. This leads to a pairwise alignment where the minimal distance score $d_m(deu, nld) = min(d(bErk - bErx))$ is identical to the distance score obtained from the pairwise alignment via PMI-based NW. The overall distance score obtained from the analysis is included in the distance matrix computation. A concept tree with an identical set of leaves as the language tree can be reconstructed, i.e. each language is present once at the leaves of the tree.
This approach takes all possible words for one meaning into account and determines a distance score depending on the languages under comparison. Since the overall

distance score is computed using one of all possible alignments, the minimal distance value is a fair measurement which is consistent with the distance computation between two languages containing no synonyms. The method represents the distance between two languages containing synonyms in a reasonable way, compared to the distance computed using randomly chosen words. The minimal distance score represents the highest similarity between two languages, indicating their closeness in the language clustering, which is represented in the tree. A possible effect of the synonyms on the language relationships can be included in the analysis with respect to the condition that a language can only occur once in the concept tree.

4.1.2 Cognate Data and Distance Matrix Combinations

The reconstruction of concept trees using the PMI-based NW alignment takes shared patterns of words into account. However, no information on cognacy is included in the analysis. Cognates are etymologically related, since they share a common ancestor. This can shed light upon both word relationships within the same language family, indicating inheritance and relatedness between words due to contact indicating horizontal transmission. Within the classical comparative method, the identification of sound correspondences and cognate classes is a circular process, which can lead to the identification of both scenarios. The integration of cognate class information is therefore a reasonable parameter which is added to the alignment analysis. Willems et al. (2016) proposed an approach to combine cognate classes and sequence alignments to compute language distances and generate distance matrices. They used the linguistic database introduced in Dyen et al. (1992), which contains manually annotated cognate judgments. Since NELex does not contain expert cognate judgments, the cognate classes are inferred automatically. For this task, the OnlinePMI program by Rama et al. (2017) is used. The program combines the PMI-based alignment method with the *Infomap* clustering algorithm originally introduced by Newman and Girvan (2004). The words within a concept are clustered into cognate classes, as illustrated in table 4.1. To compute the distance between two languages in one concept, the cognate class information is combined with the distances received from the sequence alignment:

$$d_m(l_1, l_2) = \frac{\sum_{c \in C_m} d_c(l_1, l_2)}{n_{l_1, l_2}} \tag{4.5}$$

Let $d_m(l_1, l_2)$ be the distance for meaning m between the two languages l_1 and l_2. The cognate class c is member of C_m, the set of cognate classes for meaning m. The distance $d_c(l_1, l_2)$ is the distance for the cognate class, where $d_c(l_1, l_2) = 0$ if neither word form is present in c, $d_c(l_1, l_2) = 1$ if either one or the other word form is present in c, and $d_c(l_1, l_2) = d(w_1, w_2)$ is the distance computed between two words w_1 and w_2 where both word forms are present in c.[7] If the two languages under comparison contain synonyms, the same procedure to handle synonyms is used as explained above and formalized in equation 4.4. The integer n_{l_1,l_2} is the number of cognate classes present in m that included at least one word of either l_1 or l_2.

Given the same language pairs from the previous example, a computation can be carried out using arbitrary distances for illustration purposes and the manually assigned cognate classes in table 4.1. First, the distance between *Swedish* and *Norwegian* for the meaning 'mountain' is computed. Both languages contain synonyms. The synonyms, however, belong to different cognate classes, i.e. the comparison of the lexical items belonging to the same cognate class is automatically integrated in the analysis. This is reflected in the computation, where for c in C_m the sum of all distances is $(d_m = 0) + (d_f = 0.1) + (d_b = 0.1) = 0.2$. Since the words from both languages are present in two cognate classes, $\frac{0.2}{2}$ results in the overall distance $d_{mountain}(swe, nor) = 0.1$. In the second example, the sum of all distances between *Swedish* and *German* is $(d_m = 0) + (d_f = 1) + (d_b = 0.1) = 1.1$. Swedish contains a synonym clustered in a different cognate class, i.e. the value for the distance d_f in this cognate class is set to 1. This affects the computation of the overall distance $d_{mountain}(swe, deu) = \frac{1.1}{2} = 0.55$. In the third example, the words for *German* and *Dutch* belong to the same cognate class and no synonyms are present in both languages. The sum of all distances for the cognate classes is $(d_m = 0) + (d_f = 0) + (d_b = 0.1) = 0.1$ and the overall distance results in $d_{mountain}(deu, nld) = \frac{0.1}{1} = 0.1$. The three examples illustrate the effect of cognate information combined with sequence analysis on the computation of the distances. The impact on the distance matrices depends on the cognate clustering, and therefore on the automatic cognate clustering algorithm. This, of course, can influence the reconstruction of the concept tree and its language clustering.

The approach illustrated in equation 4.5 is used to compute a distance matrix for each concept in NELex using pairwise word alignment and the corresponding automatically inferred cognate classes. Willems et al. (2016) computed the distance between two words in $d_c(l_1, l_2)$ via the normalized Levenshtein distance. In the state-of-the-art string alignment used so far in CHL, the weighted NW algorithm improves the distance measure compared to the Levenshtein method (Jäger, 2013b).

[7] For a detailed description of the formula, please refer to Willems et al. (2016).

However, a similar analysis to compare the two distance measures was not carried out within this framework, i.e. both distance measurements are taken into account in the following analyses. In the first analysis, $d(w_1, w_2)$ is computed via the normalized Levenshtein distance, as originally proposed by Willems et al. (2016). The second analysis uses the PMI-based NW algorithm to compute the distance between two words $d(w_1, w_2)$. The resulting similarities are converted into distances using d_{sig} (equation 4.2). In accordance with Rama et al. (2017), d_{sig} is the method of choice, since it is used in the OnlinePMI algorithm to convert the similarities of the sequence alignments needed for the cognate clustering. Both methods result in distance matrices for all concepts, which serve as input for the tree reconstruction methods.

4.1.3 Combining Word Distances with Geographical Distances

In a classical cross-linguistic study carried out with the classical comparative method, language relationships between geographically close languages are implicitly considered in the manual language comparison task. Geographically close languages are intuitively assumed to be related, and are first considered for a comparison to determine their relationships. The geographical spread of most of the language families, e.g. Germanic and Romance, underlines this assumption. In addition, the nodes in the tree model can correspond to both a language and its social community. Each split of a mother language into its daughter languages can therefore be seen as an event of social split into separate social communities (François, 2015).[8] A division of a population into two social communities indicates geographical closeness, which can also be assumed for the corresponding languages. This is reflected in the clustering of a phylogenetic tree, where the evolution of languages and their communities implies that genetically related languages are in most cases geographically close to each other.

Several linguistic studies which confirm this view adapted methods proposed in the field of *phylogeography*. Phylogeography combines the disciplines of biogeography, genetics, and phylogenetics to model the geographic distribution of individuals, e.g. population expansion and migration. The adapted methods can be used to combine different parameters within one analysis. In linguistics, methods are adapted to model language expansions, or to detect geographical areas of the languages and their homeland. Hereafter, the most relevant studies which served as inspiration

[8] The development and division of social communities is by far not so straightforward as assumed by the classical tree model (François, 2015). The history of populations is complex and studied in the field of sociolinguistics.

and basis for the following analyses are summarized. A method to determine the homeland of a language family was developed by Wichmann et al. (2010). In their study, measurements for geographical and lexical distance are combined to locate the language with the highest diversity in the sample. This language is then identified as the homeland of the language family.[9] The study of Atkinson (2011) is based on the assumption that languages expanded out of Africa. He claimed that the phoneme inventory size of a language decreases if the geographical distance to Africa increases. Wichmann et al. (2011) refined this approach by taking an inferred homeland, derived by the method introduced in Wichmann et al. (2010), to measure the correlation of phoneme inventory size with geographical distance to the homeland.[10]

The assumption that genetically related languages are geographically close to each other, and the idea of phylogeography to combine various distance measures, results in a newly developed method for matrix computations. The decision basis of the tree reconstruction algorithms are the distances between languages, which are represented in a matrix. It can be expected that additional parameters, integrated into the computation of the distance matrices, can improve the results of the language clustering. The geographical distance between two languages adds a stable component to the alignment analyses to take geographical closeness into account. The general idea is to add a smaller distance to adjacent languages and a larger distance to distant ones, i.e. the geographical distance serves as global parameter in the analysis. In terms of concept tree reconstruction, the impact of short sequence alignments on the language clustering can only be estimated. The number of sounds in a phonological representation of a lexical item determines the number of sounds which can be aligned by the algorithms. This can have an effect on the computed distances, i.e. short words can lead to small distance variations in the concept distance matrix. Since the tree reconstruction algorithms highly depend on the distance matrices, small distance variations can affect the resulting language clustering. The geographical language distances add a stable component to the analysis, which can improve the language clustering on the concept tree in order to get a clearer insight into the language relationships.

The geographical distances between language pairs are computed using the method introduced in Wichmann et al. (2011). Since only the computation of the geographical distances between language pairs is needed and not the inferred homeland, the

[9] In another prominent study, Bouckaert et al. (2012) used Bayesian phylogeographic approaches combined with vocabulary lists to trace back the homeland of the Indo-European language family.

[10] The phylogeographic program of Wichmann et al. (2011) is available on the homepage of the ASJP database (Wichman et al., 2018).

corresponding code is extracted and modified.[11] For each language pair in NELex, the geographical distance is computed using the longitudes and latitudes for each language.[12] A geographical distance matrix is generated, having the same form than a concept distance matrix (see e.g. the concept matrix in table 3.5). This means the entries along the diagonal represent the geographical distances of same languages, which are 0, while the off-diagonal entries represent the geographical distance between different languages. Each concept distance matrix can therefore be merged with the geographical distance matrix. The merging of the two matrices is done via the *super distance matrix* (SDM) method proposed by Criscuolo et al. (2006), which is implemented in the R package *ape* (Paradis et al., 2004). The SDM method fuses both distance matrices into a supermatrix, which can be used to reconstruct a concept tree. To merge two matrices, "SDM deforms the source matrices, without modifying their topological message, to bring them as close as possible to each other; these deformed matrices are then averaged to obtain the distance supermatrix" (Criscuolo et al., 2006, p. 740). The SDM algorithm is described in detail in Criscuolo et al. (2006), where the results of simulation studies show that "SDM deals efficiently and accurately with collections containing a large number of source matrices of varying size" (Criscuolo et al., 2006, p. 741). The SDM method reconstructs a supermatrix that is the best representative of all input matrices, i.e. one concept distance matrix and the geographical matrix. Since the shape of the distance matrix remains the same in the supermatrix, it can serve as input to the distance-based tree reconstruction algorithms.

4.1.4 Reliability Measure for Distance-Based Concept Trees

The application and development of distance computations for the reconstruction of concept trees using phylogenetic methods is a feasible task in CHL. However, an evaluation of the automatically reconstructed trees is needed to get an idea of the suitability of the algorithms in CHL. There are two techniques to determine the reliability of trees.

Correctness of Trees The best attested technique to verify the correctness of an automatically reconstructed tree is the comparison to an expert tree, i.e. a tree reconstructed by linguists (Jäger, 2013a). As already mentioned in chapter 3.1.3, there are several databases like *Glottolog* (Hammarström et al., 2018) providing

[11] For a detailed description, please refer to the source code provided by the ASJP database (Wichman et al., 2018).

[12] The longitudes and latitudes for the languages are included in NELex (Dellert, Jäger, et al., 2017).

expert classifications of linguistic trees which can be used to evaluate the automatically reconstructed trees obtained from the algorithms. To measure the correctness of distance-based concept trees is, however, not straightforward. The expert tree from Glottolog displays the history and relationships between languages, whereas a concept tree represents relationships between lexical items of a specific concept for these languages. Since the representation purposes of the trees differ, the expert tree from Glottolog cannot serve as gold standard tree for a comparison with concept trees. The representation of word histories using trees is a new approach, and no expert words trees are available which could be used for a tree comparison. To measure the correctness of concept trees is therefore unfeasible.

Stability of Trees The stability of the tree and its clades is determined using statistical resampling techniques. Resampling techniques are used to estimate the precision of sample statistics. The procedure is as follows. First, a new dataset is created by either using a subset of the initial dataset, or by choosing randomly with replacement characters from the initial dataset to create a sample of the same size. Second, the analysis is replicated using the sample sets in order to estimate sample statistics, like e.g. medians and variances. The two standard resampling techniques are *jackknife* and *bootstrap*. The jackknife creates subsets from the initial dataset by dropping one or more characters and calculating the estimate each time (Felsenstein, 2004). The bootstrap involves sampling with replacement to create a sample set of the same size as the initial dataset, which is used to rerun the analysis. To illustrate the general ideas of the two techniques, the language sample displayed in table 3.1 is used, which is repeated in table 4.2 below.

	Meaning 'sun'		Meaning 'mountain'		Meaning 'you'		Meaning 'louse'		
	word	cc	word	cc	word	cc	word	cc	
English	son	s	mauntin	m	yu	d	laus	*l*	...
Icelandic	soul	s	fEtl	f	8u	d	lus	l	...
Swedish	sul	s	bErg, fyEl	b,f	d3	d	l3s	l	...
Norwegian	sul	s	bErg, fyEl	b,f	d3	d	l3s	l	...
Dutch	zon	s	bErx	b	y3	d	l3is	l	...
German	zon3	s	bErk	b	du	d	laus	l	...
Romanian	soare	s	munte	m	tu	d	p3duke	p	...
Italian	sole	s	monta5a	m	tu	d	pidokyo	p	...
French	solEy	s	mota5	m	ti	d	pu	p	...
Catalan	soL	s	munta53	m	tu	d	pol	p	...
Portuguese	soL	s	mota5a	m	tu	d	pyulu	p	...
Spanish	sol	s	monta5a	m	tu	d	pyoxo	p	...
...									

Tab. 4.2.: An excerpt of the data from the NorthEuraLex Database containing the phonological representations of the words in the ASJP alphabet and manually assigned cognate classes.

If $s = \{sun, mountain, you, louse\}$ is the set of all concepts, the data is sampled depending on the method of choice to create different samples. The jackknife would create subsets of a smaller size without replacement, i.e. each concept can only be present once, e.g. $s_1 = \{sun, mountain, you\}$ and $s_2 = \{mountain, you, louse\}$. Bootstrap would sample with replacement, i.e. a concept can appear more than once in a set in order to create k sample sets of the same size, e.g. if $k = 2$ then $s_1 = \{sun, sun, mountain, louse\}$ and $s_2 = \{mountain, louse, you, you\}$. For each sample in the set, a tree is reconstructed generating a set of trees. This would result in two trees for both sampling techniques. For each method, the resulting trees are summarized in a *consensus tree*, where the stability values of the inner clades reflect the appearance of the corresponding clades in the tree sample. The two main methods to reconstruct a consensus tree are the *strict consensus tree* and the *majority-rule consensus tree*. The strict consensus tree contains all clades present in the tree sample, which can lead to an unresolved consensus tree (Felsenstein, 2004). The majority-rule consensus tree represents the clades that are present in the majority of the trees, i.e. typically in over 50% of the trees in the sample.

Standard Method: Jackknife

The jackknife technique (Felsenstein, 1985; Mueller and Ayala, 1982) is the older technique and samples without replacement. The two main techniques are the *delete-half* and the *delete-one* jackknife. The underlying concepts are illustrated using the data in table 4.2, where the set of concepts is of size $n = 4$. The delete-half jackknife samples $\frac{n}{2}$ times without replacement, i.e. a random half of the characters are taken into account. This results in two sample sets of size $n = 2$, e.g. $s_1 = \{sun, mountain\}$ and $s_2 = \{you, louse\}$. The delete-one jackknife drops one observation at a time, i.e. each sample includes $n - 1$ concepts. This results in four sample sets of size $n = 3$ with the following data: $s_1 = \{sun, mountain, you\}$, $s_2 = \{sun, you, louse\}$, $s_3 = \{sun, mountain, louse\}$, and $s_4 = \{mountain, you, louse\}$. In linguistics, it is advisable to use the delete-one jackknife, since this technique results in a larger sample size. This is relevant due to the sometimes small datasets used for linguistic tree reconstruction.
In the case of language tree reconstruction, the delete-one jackknife results in a number of matrices of the same size $n - 1$. Language trees are reconstructed using all concepts present in the database, which means the number of concepts determines the sample size. Since the number of concepts included in the digitally available databases varies between 40 and 1,016, sampling without replacement is unproblematic. However, in terms of concept tree reconstruction, the application of

the delete-one jackknife is problematic. A concept sample normally contains one lexical item per language for the corresponding concept. In the case of synonyms, the number of lexical items might increase slightly, i.e. in most cases there is one additional synonym present. The columns of the matrix determine the sample size of the jackknife method. The concept data displayed in table 4.1 shows that without synonyms, the data is represented in a vector, and with synonyms in a two-column matrix. Dropping one observation at a time would result in a pretty small sample set in case of a two-column matrix, whereby it is unfeasible for a vector. An alternative approach would be to apply the delete-one jackknife on the rows of the matrices, i.e. dropping one language at a time. Felsenstein (2004, p. 357) states that "it is not easy to see what statistical meaning this jackknifing of [languages] will have. [Languages] are not independent and identically distributed - they come to us on some phylogeny, where they are highly clustered." Additionally, jackknifing the rows of the matrices would result in a matrix containing $n-1$ languages from the sample, i.e. the set of leaves for each reconstructed tree contain different languages. The trees in the sample can therefore not be summarized in a consensus tree, since for this reconstruction all trees need to have the same set of leaves. However, a number of sample trees with different sets of leaves can be summarized using a *supertree*. Supertree techniques are divided into direct and indirect approaches. Direct methods reconstruct a supertree out of a tree sample without providing any support information on the inner clades of the tree to indicate the stability of the tree. Indirect methods generate a matrix out of the tree sample which serves as input for the supertree reconstruction. In the resulting supertree, only goodness-of-fit statistics, like e.g. consistency values for the inner clades, can be displayed (Bininda-Emonds et al., 2002). This is not the same as computing support values to determine the stability of the inner clades to reflect the clustering of the trees in the sample. Both direct and indirect supertree methods can therefore not be used to estimate the stability of concept trees. The underlying concept of the indirect supertree technique leads to the idea of using the SDM method to combine various matrices in order to reconstruct a supertree. By applying the delete-one jackknife procedure, $n-1$ jackknife matrices are generated by leaving out one language at a time from the sample with n languages. The distance matrix computation is repeated for each sample, creating a set of matrices. The distance matrices of all jackknife replicates can be combined using SDM, out of which a single tree is reconstructed. Since no tree sample can be created out of this method, there is no possibility to compute the frequency of the clades in order to reflect their stability. To conclude, the delete-one jackknife cannot be applied to concept data in order to assess clade credibility to concept trees.

Standard Method: Bootstrap

The bootstrap technique, introduced by Efron (1979), involves sampling with replacement to create k number of bootstrap replicates. The variability of the estimate is inferred by the bootstrap replicates, using a rather large variation of the initial sample (Felsenstein, 2004). The size of the bootstrap sample k is set manually. Since the bootstrap technique samples with replacement, an appropriate value for k needs to be estimated in order to obtain a large variation of the initial sample. Typically, the sample size of bootstrap replicates varies between 100 and 1,000. The result is a sample of k data matrices with the same number of languages and the same number of concepts as in the initial data matrix of size n.

Since the words are sampled with replacement, one lexical item can occur more than once in the sample, whereas another one might not occur at all. In the case of language trees, bootstrapping can unproblematically be applied to the basic vocabulary lists. Using the data in table 4.2, the set of concepts is of size $n = 4$, and $k = 3$ is the manually set sample size of the bootstrap replicates. Three datasets of size $n = 4$ are replicated by randomly sampling with replacement, resulting e.g. in $s_1 = \{sun, sun, mountain, louse\}$, $s_2 = \{mountain, louse, you, you\}$, and $s_3 = \{mountain, mountain, you, you\}$. For each dataset in the sample, the analysis is repeated following the initial procedure for matrix computation and tree reconstruction. The resulting tree sample of bootstrap replicates is also of size k. All sample trees are summarized in a majority-rule consensus tree, including only those clades which are present in the majority (50%) of the trees. An estimate for the level of support on each clade is provided, determining the stability of the clusters in the trees. Compared to the jackknife technique, the bootstrap method is independent of the number of characters in the data sample, since the number of replicates is set manually and can be customized for each analysis. Therefore, the bootstrap method has become the standard technique for the assessment of clade credibility in phylogenetic trees in phylogenetics as well as in linguistics.

On the concept level, the traditional bootstrap method leads to the question how to sample from a word list with replacement. As shown in data illustrated in table 4.2, most concepts include one representation per language, and only a few synonyms might be present in the dataset. Sampling the columns of the matrix with replacement would lead to a sample of size k with identical matrices. The sample trees would be highly similar, if not identical. This results in a majority-rule consensus tree, which would not be a good estimate for clade stability and tree credibility as expected from a bootstrap analysis.

A logical conclusion would be the sampling of phonetic characters instead of lexical items. From a theoretical point of view, this would be possible. The characters

can be sampled with replacement in order to create bootstrap replicates which can serve as input for the matrix computation and tree reconstruction algorithms. However, the phonetic characters of a lexical item appear in an order which reflects the phonological system of the language and follows certain word formation rules. The order of the sounds has a significant meaning which should not be split by a sampling method. The rearrangement of the sounds would result in nonsense words. This has a great impact on the sequence comparison and the following analysis, i.e. the concept trees would no longer illustrate the history of a concept, and the majority-rule consensus tree would not be a good estimate for clade credibility.

Noisy Bootstrap

The jackknife and bootstrap techniques cannot be applied to achieve the intended results for measuring the stability of distance-based concept trees. An alternate method, developed by Nerbonne et al. (2008), can be transferred for this task. Nerbonne et al. (2008) introduce a "noisy" clustering approach to overcome instabilities in dialect clustering. The fundamental idea of this technique is to add noise to a distance matrix to indicate whether the signal or the clade in a tree is stable within different replicates. This technique is adapted to measure the stability of the clades in concept trees where sampling of the data is impracticable.
The stability of the inner clades of a tree is measured by modifying the concept matrix in order to obtain varying replicates which can be summarized in a consensus tree. Following Nerbonne et al. (2008), a small noise ceiling c, where $c = \sigma/2$, i.e. one half of the standard deviation of the distances in the corresponding matrix, is specified. Random amount of noise r, where r varies uniformly ($0 \leq r \leq c$), is added to the matrix to generate a new matrix replicate out of which a tree is reconstructed. The computation is done $k = 1,000$ times to create a tree sample of size 1,000.[13] A majority-rule consensus tree summarizes the tree sample to display the support values of the inner clades. Since the procedure is similar to the bootstrap technique, the method will be called *noisy bootstrap*.

Evidence for Noisy Bootstrapping An experiment should shed light on the accuracy and efficiency of the noisy bootstrap method. For this task, a language sample as displayed in table 4.2 serves as input to reconstruct a distance-based language tree. Both the standard and the noisy bootstrap technique can be applied to determine the stability of the tree clustering. The stability values obtained from both methods can

[13] The size of k should be higher than 100 because the variance of the matrices is small. A sample of size $k = 1,000$ can give clearer insights into the stability of the inner clades.

be compared to evaluate the accuracy of the noisy bootstrap method. Additionally, the automatically inferred tree and both consensus trees can be compared to an expert tree to account for correctness.
For this task, a subset of the NorthEuraLex database is extracted, including all Indo-European languages with the most 200 stable concepts (Dellert and Buch, 2018).
To determine the correctness of reconstructed trees, an expert tree is extracted from the Glottolog database (Hammarström et al., 2018) for the comparison. The tree displays the expert classifications of the Indo-European languages included in the subsample from NELex and serves as gold standard tree for the comparison. Two trees are compared by computing the topological distance between them. The smaller the distance, the better the automatically reconstructed tree. In linguistics, language trees can either be bifurcating or multifurcating. Distance-based trees are bifurcating, whereas expert and consensus trees can be multifurcating. Following Jäger (2013a) and Jäger (2013b), the *generalized quartet distance* (GQD) is used as distance measure between two trees. It is a generalized version of the quartet distance introduced by Estabrook et al. (1985), which takes the asymmetry between bifurcating and multifurcating trees into account. The method compares existing quartets in the expert tree to the quartets in the inferred tree, and computes the distances as the proportion of quartets on the leaves, which have a different topology.
The underlying procedure for language tree reconstruction is introduced in chapter 3.1.3. To determine the distance between two languages, the wordpair distances are computed using the PMI-based NW algorithm, and aggregated using the measurement called *dERC* (Distance-based on Corrected Evidence of Relatedness) in Jäger (2013b).[14] The resulting distance matrix serves as input to the FastME algorithm to reconstruct the language tree. The standard bootstrap method can be applied to language data. The lexical items in the dataset are sampled with replacement to generate k new data matrices, where $k = 1,000$. For each data matrix, the initial analysis, including matrix computation and tree reconstruction, is repeated. The tree sample contains 1,000 replicates, out of which a majority-rule consensus tree is reconstructed. For the noisy bootstrap technique, the distance matrix computed for the language data is used to add noise and create $k = 1,000$ noisy bootstrap matrices. Tree sample of size 1,000 is used to compute the corresponding majority-rule consensus tree. In terms of efficiency, it can already be seen that adding noise to a matrix is computationally cheaper than the procedure of standard bootstrap.
All three output trees, i.e. the language tree and the two consensus trees, are compared to the expert tree of Glottolog using the GQD. The results are shown in table 4.3.

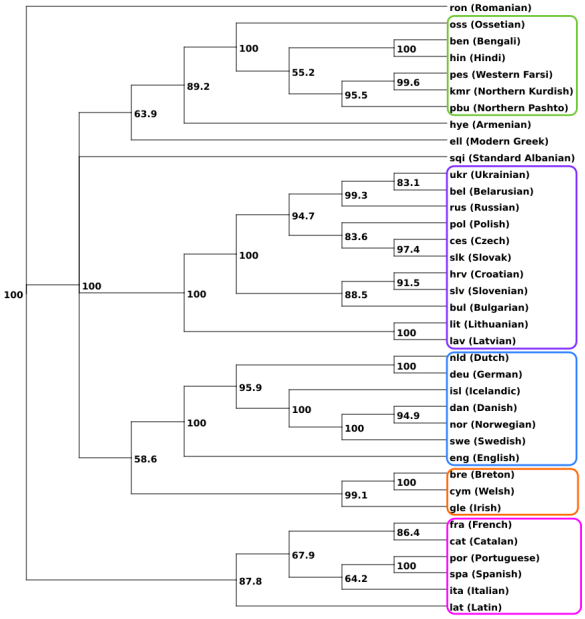
[14] For a detailed mathematical description of the dERC method and the procedure of aggregating language distances, see Jäger (2013b).

Tree compared to Glottolog tree	GQD
Inferred language tree	0.0193
Consensus tree standard bootstrap	0.0193
Consensus tree noisy bootstrap	0.0206

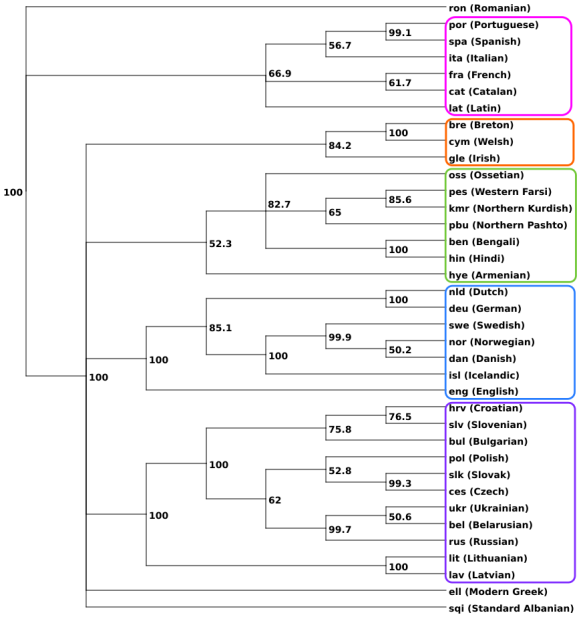
Tab. 4.3.: Generalized quartet distance between the three inferred trees and the expert tree

All three trees show a high agreement with the expert classification, i.e. the GQD distances are small. The distance between the automatically reconstructed language tree and Glottolog indicates a correctness of the inferred tree from over 80%. This is in accordance with the results obtained in other linguistic studies, see e.g. Steiner et al. (2011), Jäger (2013a), and Jäger (2013b). The small distance between the consensus tree generated out of the tree sample from the standard bootstrap method and the expert tree indicates that data sampling results in a stable tree. The distance between the expert tree and the consensus tree computed from the noisy bootstrap method is slightly, but not significantly, larger. The consensus tree still shows a high agreement around 80% to the Glottolog tree, which underlines the correctness of the consensus tree. This leads to the result that the noisy bootstrap method is a suitable alternative to the standard bootstrap technique in terms of accuracy.[15]
The stability values obtained from the standard and the noisy bootstrap techniques can be compared to get a clearer insight on the accuracy of the clustering. The trees are displayed in figure 4.1. Both majority-rule consensus trees are unrooted and multifurcating, which is mirrored in their topologies. The topologies differ slightly in their internal clustering, which is due to the different computation techniques of the replicated distance matrices and reflected in the GQD distances. The languages are clustered correctly together according to their families. Within the language families, the internal structure is similar in both trees. The stability values are higher on the standard bootstrap consensus tree, indicating a greater stability within the replicates. It is not surprising that a reiteration of the initial analysis using data sampling leads to more stable results than adding noise to the distances. However, the slight difference between the stability values does not have a great impact on the accuracy of the methods. In most cases, there will be variance within the automatically reconstructed (consensus) trees, which is due to the various computational approaches. Nevertheless, the comparison of the consensus trees supports the noisy bootstrap method as an alternative technique to standard bootstrap.
The experiment verifies the assumption that noisy bootstrapping can be used to determine the stability of trees where statistical sampling is impracticable. In terms of efficiency, the noisy bootstrap method is computationally cheaper. Compared

[15] For the sake of completeness, all analyses are also computed using the Neighbor-Joining algorithm. For a comparison of the two reconstruction algorithms, see Appendix A.1.



(a) Consensus tree standard bootstrap.



(b) Consensus tree noisy bootstrap.

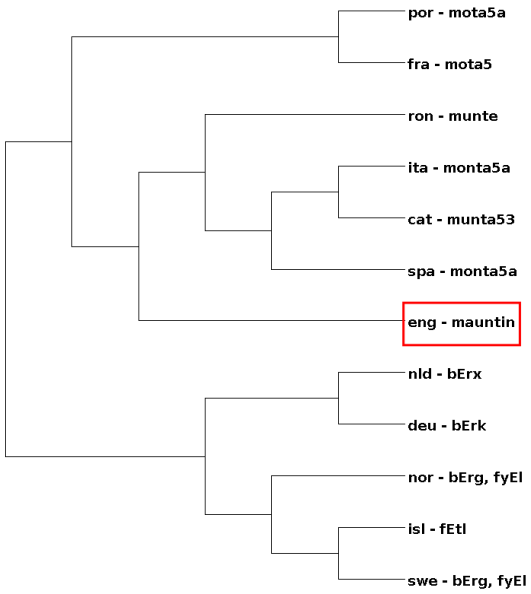
Fig. 4.1.: Graphical representation of the two consensus trees from the bootstrap analyses. Language families are colored: pink=Romance, blue=Germanic, green=Iranian, purple=Slavic, orange=Celtic.

to data sampling in order to repeat the initial analysis, adding noise to a distance matrix speeds up the analysis and results in a faster running time. The comparison of the noisy bootstrap consensus tree to the Glottolog tree, and the comparison of the stability values between the consensus trees show a high accuracy of the noisy

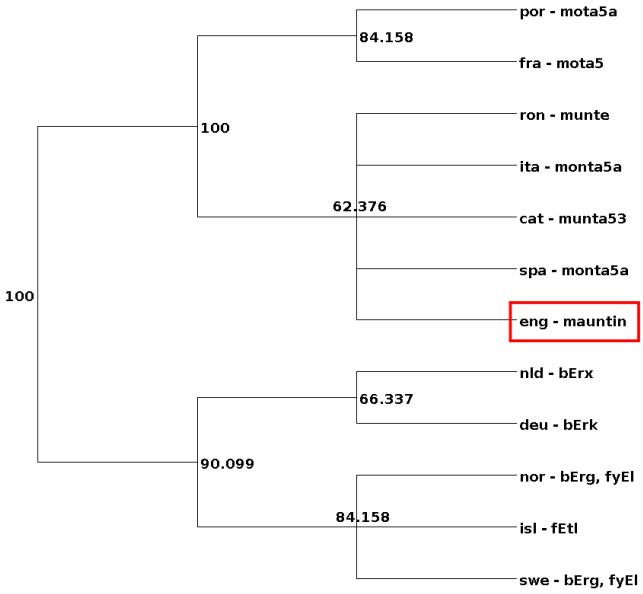
bootstrap method. Since the assessment of stability values works well for language data, the noisy bootstrap method can be transferred to concept data to get an insight into the concept trees' stability.

Noisy Bootstrapping for Concept Trees To determine the stability value of concept trees, the procedure of the noisy bootstrap method will remain. In table 3.5, a sketch of a concept distance matrix is given. In two computation steps, the noisy bootstrap method generates bootstrap matrices of size k for each concept. In the first step, the noise ceiling c, where $c = \sigma/2$, is specified. Second, random amount of noise r, where r varies uniformly ($0 \leq r \leq c$), is added to the matrix. These two steps are repeated $k = 1,000$ times, and a tree sample is generated. The trees are summarized using a majority-rule consensus tree. The concept tree for 'mountain', displayed in figure 3.3, is repeated in figure 4.2a for a direct comparison with the consensus tree reconstructed from the noisy bootstrap method. The consensus tree in figure 4.2b shows unresolved clusters within the Germanic and the Romance language family. This illustrates an unstable clustering within the bootstrap replicates, where none of the possible clustering combinations were present in more than 50% of the bootstrap trees. The stability value reflects the frequency of appearance of this cluster within the bootstrap replicates. The first cluster of the Romance language family, which includes English, has a stability value of 62%. The other cluster, including French and Portuguese, has a stability of 84%. Although the Romance languages descend from Latin, they additionally consist of loanwords from Latin which were later integrated into the languages. Latin is the most common source of loanwords in the Romance languages, in addition to a great amount of internal borrowings between the Romance languages (McMahon and McMahon, 2005). This uncertainty is reflected in the clustering of the consensus tree, and is even more present in the automatically reconstructed concept trees. The Germanic languages are divided into two clusters. The first cluster represents the West Germanic language family, including German and Dutch, with a stability of 66%. The North Germanic languages are clustered together with a stability of 85%. The high stability values of the Germanic and Romance languages show a clear and stable distinction between the two language families. In terms of loanword detection, the grouping of English within the Romance languages indicates a stable position of English within the Romance languages in the replicated trees. The transcription of the words for the concept 'mountain' already indicates the high similarity to the Romance languages, which is confirmed in the clustering of the consensus tree.

The noisy bootstrap method can be applied to concept data to determine the stability of the tree clustering. All possible reconstruction methods introduced so far have an



(a) Distance-based concept tree.



(b) Consensus tree noisy bootstrap.

Fig. 4.2.: Graphical representation of the distance-based concept tree and the consensus tree of the noisy bootstrap method for the concept 'mountain' using an excerpt of the Germanic and Romance languages.

underlying distance matrix for each concept where the noisy bootstrap technique can be applied.

4.1.5 Evaluation

The different matrix computation methods introduced so far result in more than one reconstruction option. In terms of automatic loanword detection, the method with the most representative trees for the concepts should be determined. On the one hand, tree reconciliation using all trees reconstructed from the introduced methods is unfeasible. On the other hand, the optimal method is assumed to lead to more accurate results in terms of loanword identification, since the matrix computations lead to the most representative trees. Several approaches are worth considering to evaluate the different methods for matrix computation.

Species Tree Reconstruction This approach is adapted from phylogenetics, where a species tree can be reconstructed out of different gene trees. In linguistics, the concept trees can be summarized by a consensus method to reconstruct a language tree. The distance between the language consensus tree and the expert Glottolog tree can be computed to measure the agreement of the two trees, i.e. the correctness of the language consensus tree. However, due to missing entries in NELex, the concept trees do not contain the same set of leaves, i.e. no consensus tree can be reconstructed, and therefore no comparison to an expert tree can be made.

Consensus Tree Reconstruction Another approach is the usage of the consensus tree for each method, reconstructed out of the noisy bootstrap tree replicates. Each concept consensus tree would contain stability values at the inner clades, which could be compared between the concepts of the different methods. This would imply that trees have the same set of inner nodes. However, the concept trees will only have the same topology by chance, since the reconstruction of different replicates will lead to different concept consensus trees for each method.

MCC Tree Reconstruction An alternative to a consensus tree is the *maximum clade credibility* (MCC) tree, introduced by Heled and Bouckaert (2013). The maximum clade credibility method evaluates each tree in the sample to find the best representative. In contrast to the consensus tree, the MCC tree is an actual tree from the sample of replicates, and not a summary of the inner clades. In addition, the MCC tree is a fully resolved topology, whereas a consensus method might produce unresolved trees (Heled and Bouckaert, 2013).
The maximum clade credibility method computes a score for each clade in a tree on the fraction of times this clade appears in the trees of the sample. The product

of the score is taken as the *clade credibility score* for the tree. The maximum clade credibility tree is the tree with the highest clade credibility score, and therefore the best representative tree of the tree sample.[16] For each concept, the highest clade credibility score is computed using the tree replicates obtained from the noisy bootstrap method. The clade credibility scores between all methods are compared in order to identify the best representative tree, i.e. the tree with the highest clade credibility score. The method with the highest sum of highest clade credibility scores for all concepts serves as the ideal method for further analyses.[17]

In the previous chapters, various matrix computation methods for automatic tree reconstructions were introduced. In total, there are eight different possibilities to compute concept distance matrices and reconstruct trees using FastME, as listed in table 4.4.

Method	Description
CT_{dow}	PMI-based NW algorithm + d_{dow} conversion
CT_{sig}	PMI-based NW algorithm + d_{sig} conversion
$CT_{dow+Geo}$	PMI-based NW algorithm + d_{dow} conversion + geographical distances
$CT_{sig+Geo}$	PMI-based NW algorithm + d_{sig} conversion + geographical distances
CT_{Lev}	Levenshtein distance + cognate classes
CT_{NW}	PMI-based NW algorithm + d_{sig} conversion + cognate classes
$CT_{Lev+Geo}$	Levenshtein distance + cognate classes + geographical distances
CT_{NW+Geo}	PMI-based NW algorithm + d_{sig} conversion + cognate classes + geographical distances

Tab. 4.4.: The eight different matrix computation methods for tree reconstructions with FastME.

The first two computations (CT_{dow} and CT_{sig}) use the PMI-based NW algorithm and convert the similarities into distances using two different measurements, namely d_{dow}, introduced in equation 4.1, and d_{sig}, stated in equation 4.2. Two approaches (CT_{Lev} and CT_{NW}) use the idea introduced in chapter 4.1.2, the combination of cognate class information with a distance measure for lexical items. One approach uses a normalized Levenshtein distance in correspondence with Willems et al. (2016),

[16] See Heled and Bouckaert (2013) for mathematical details on the computation of the clade credibility score.

[17] The computation of the maximum clade credibility score for a tree sample is implemented in the python package *Dendropy* (Sukumaran and Holder, 2010).

who introduced this idea. The other method uses the PMI-based NW algorithm with the sigmoid conversion to transform similarities into distances. All resulting distance matrices from the four approaches can be combined with the geographical distance matrix computed for the languages in NELex. For each concept in the eight methods, the noisy bootstrap method is used to create a tree sample of 1,000 replicates.
The ideal method is determined using the concept of an MCC tree and the clade credibility computation. For each method, the best representative tree is found for each concept out of 1,000 noisy bootstrap replicates. The method with the highest sum of highest clade credibility scores for all concepts serves as optimal method. The methods and their corresponding sum of the highest clade credibility scores are listed in table 4.5.

Method	SCC
CT_{dow}	**503**
CT_{sig}	14
$CT_{dow+Geo}$	205
$CT_{sig+Geo}$	143
CT_{Lev}	0
CT_{NW}	0
$CT_{Lev+Geo}$	33
CT_{NW+Geo}	118

Tab. 4.5.: The MCC comparison of the different matrix computation approaches for the tree reconstruction with FastME. The ideal method with the sum of highest clade credibility scores for the concepts is marked in bold.

According to the sum of highest clade credibility scores for all concepts, the CT_{dow} method serves as the ideal method for distance-based concept tree reconstruction. The CT_{dow} method uses the equation 4.1 to convert the similarity scores of the PMI-based NW algorithm into distances. This indicates that in the case of concept trees, the conversion method introduction by Downey et al. (2008) results in a better distance estimation than the sigmoid function used by Rama et al. (2017). The sigmoid function transforms the similarities into distances of range $[0, 1]$, i.e. it scales the distances in order to fit into the range. The scaling of the distances into a small range brings them closer together, as expected by the similarity score obtained from the PMI-based NW algorithm. Since the distances in the matrix are the decision basis of the tree reconstruction algorithm, the scaling of the sigmoid function affects the language clustering and could result in unsuitable trees. Out of the four methods using cognate classes plus a distance measurement for words within the same cognate class, the PMI-based NW algorithm combined with geographical distances leads to the best results. This is not surprising, since the Levenshtein distance only distinguishes between identical and non-identical sounds, whereas the PMI-

based NW algorithm achieves a better approximation of the sound similarities using PMIs and results in better alignments (Jäger, 2013b). The results of the analyses underlie this assumption. However, the introduction of cognacy as additional parameter did not add more information to the analysis, as it was expected. This can have two reasons. First, the method used to combine cognate classes with distances is not as informative as using pairwise alignments. The PMI-based NW algorithm is used to align words clustered in the same cognate class, otherwise the matrix is filled with 0s and 1s, as shown in equation 4.5. This results in distance matrices which are less informative than aligning all word pairs using the PMI-based NW algorithm. Second, the cognate classes are inferred automatically. Rama et al. (2018) showed that for language tree reconstruction on cognate data (manually and automatically inferred) compared to an expert Glottolog tree, the reconstructions using expert cognate judgments are generally more suitable. The difference to the reconstructed trees using automatic cognate sets is, however, not very large and therefore not significant. This indicates that automatically inferred cognate classes can be used to reconstruct reliable phylogenetic trees on the language level. Rama et al. (2018) reconstructed the trees using automatic cognate data in a Bayesian framework, which can lead to different results than using cognate class information in a distance-based framework. Additionally, the impact of cognate classes on the language level can be different than the effect on a concept level. This is reflected in the results, since most of the other methods outperform the ones using cognate data. Adding geographical distances to the cognate class distances improves the results. The geographical distances between languages add a stable component to the analysis to take geographical closeness into account. The parameter indicates that geographically adjacent languages are closer together in the distance matrix and therefore also in the tree, which could have an effect on the clustering of the languages. Compared to the methods using cognacy information, the geographical distance does not have a great impact in combination with pairwise sequence alignment. Using the sigmoid transformation method, the results improve by including geographical distances. However, the CT_{dow} approach still outperforms the rest of the methods, including the one combined with geographical distances.

4.2 Character-Based Methods

Character-based models of change have increasingly become a matter of interest in linguistics, since they estimate the relationship between two languages by inferring the evolution from their common ancestor. This involves the usage of an evolutionary

model according to which the evolutionary history of the characters is modeled onto a phylogenetic tree. The advantage of using models of character evolution is the description of character changes along a phylogenetic history, i.e. the phylogenetic tree optimally explains a given data matrix. The data matrices can be generated using multiple sequence alignment or binary data representations obtained from a sequence clustering, which serve as input to the algorithms of the different character-based frameworks.

The underlying procedure for the reconstruction of concept trees is introduced in chapter 3.1.4. The linguistic data needs to be organized in character matrices, where the languages are categorized according to discrete features. Ideally, the discrete features are historically informative, e.g. the clustering of words into cognate classes to represent the relationships of the languages. However, high quality data, i.e. the manual assignment of cognate classes by linguists, is only available for a few databases. The cognate classes are transformed in order to generate binary presence-absence matrices, as illustrated in table 2.1 and 4.6. In terms of concept tree reconstruction, the generation of binary matrices can lead to one difficulty, namely the assignment of few cognate classes by the automatic clustering algorithm. From a linguistic point of view, this is not surprising, since it can be the case that words are closely related within one concept. Few or even just one cognate class illustrate the descent from common ancestor(s). It is well known that words from stable concepts are related and can be inherited from a single common ancestor in one language family. This leads to the assumption that the reliability of a concept tree reconstructed with a small number of cognate sets is questionable. To gain more signal out of the characters of the lexical items for a suitable reconstruction of character matrices, innovative ideas on matrix computations are introduced in the following chapters next to the state-of-the-art methods.

In chapter 3.1.2, the different character-based approaches are introduced. The field of bioinformatics provides well-known programs for character-based analyses, which can be adapted into CHL. The maximum parsimony method aims at finding the optimal tree using the minimum number of character changes. This often results in more than one optimal tree, and it is not trivial to choose the best tree from the resulting sample. Therefore, maximum parsimony methods are excluded from concept tree reconstructions. For the frameworks of maximum likelihood and Bayesian inference, several linguistic studies (Atkinson and Gray, 2006; Greenhill et al., 2010; Gray and Atkinson, 2003; Hinneburg et al., 2007) show that the standard algorithms work well on linguistic data. In accordance with those studies, for each framework one algorithm is chosen to reconstruct concept trees.

Maximum Likelihood In chapter 3.1.2, the maximum likelihood framework is introduced. The underlying idea of maximum likelihood estimation is to find the phylogenetic tree that maximizes the likelihood under a given model of evolution. The evolutionary model specifies the evolution of the sequences along the edges of a tree, where the model assesses the probability of mutations. The lower the number of mutations, the higher the probability, i.e. the higher the likelihood. The tree with the highest probability is the optimal tree obtained from the algorithm under the given model of evolution.[18]

The main advantage of maximum likelihood estimation is the provision of "a systematic framework for explicitly incorporating assumptions and knowledge about the process that generated the given data" (Huson et al., 2010, p.40). It needs to be considered that evolutionary models are a rough estimate of the (biological) evolution known from reality. However, Huson et al. (2010, p. 40) mentioned that "in practice, maximum likelihood methods are believed to be quite robust to violations of the assumptions made in the models." This indicates that evolutionary models add important parameters to the estimation of character evolution along a phylogenetic history. On the other hand, it can be quite a challenge to determine the best evolutionary model of a given dataset for the analysis.

In linguistics, determining the best evolutionary model is not so trivial, since most of the implemented models are defined for biological data. One solution to this challenge is to choose the best model for a given dataset automatically in order to reconstruct a suitable tree under the maximum likelihood framework. The program of choice for the reconstruction of the concept trees is *IQ-Tree* (Nguyen et al., 2014). IQ-Tree provides fast tree inference algorithms to reconstruct maximum likelihood phylogenies for large datasets. Compared to other programs like *RAxML* (Stamatakis, 2014) and *PhyML* (Guindon et al., 2010), IQ-Tree computes trees with a higher likelihood in a reasonable runtime (Nguyen et al., 2014). IQ-Tree provides a modelfinder (Kalyaanamoorthy et al., 2017) to determine the best-fit model for the given data in a reasonable time. This is an advantage for concept tree reconstruction, since the model is chosen according to the data matrix given to the algorithm. The analysis is therefore independent of a manual choice by the user, which could have led to an inaccurate selection of an evolutionary model that might not be in accordance with the given data. The maximum likelihood analysis in IQ-Tree is based on an initial starting tree. The tree can be a user-defined fixed tree topology, a BIONJ tree (Gascuel, 1997) computed by IQ-Tree, or a random starting tree. The initial tree serves as basis for the application of the evolutionary model in order to estimate the character evolution on a given phylogeny to compute the maximum likelihood tree. The program contains a standard bootstrap method,

[18] A classification of the most important DNA models is given in Huson et al. (2010, p. 31).

where the number of bootstrap replicates can be specified. IQ-Tree summarizes the tree replicates using a majority-rule consensus tree, where the support values are displayed at the inner nodes of the tree. The program provides several outputs, including the reconstructed maximum likelihood tree, the consensus tree, and the tree sample containing all bootstrap replicates. In addition, IQ-Tree also returns the initial tree, a file containing the model choice, and a log file, if needed. For the following analyses, IQ-Tree is used with the following settings: a BIONJ starting tree computed by IQ-Tree, the model proposed by the modelfinder, and the standard bootstrap method with a sample size of $k = 100$.

Bayesian Inference Bayesian methods are currently the most popular methods used in linguistics to estimate trees using character-based models of change (Atkinson and Gray, 2006; Greenhill et al., 2010; Gray and Atkinson, 2003). The underlying framework is introduced in chapter 3.1.2. Bayesian inference aims at estimating the posterior probability of a phylogenetic tree via the MCMC approach. The methods result in a distribution of phylogenetic trees instead of a single tree representing a point estimate. Usually, the trees in the sample are summarized in a consensus tree. The posterior distribution provides support values of groups of taxa present in the tree sample, which are displayed at the inner nodes of the consensus tree. An overview of the work on Bayesian inference of phylogenetic trees is given by Huelsenbeck et al. (2001).

The main advantage of Bayesian methods is the computation of a tree sample according to the posterior distribution of trees, which opens up new possibilities for further processing and analyses. The posterior distribution is computed based on the input data, an evolutionary model, and a presumed prior distribution of trees. In comparison to IQ-Tree, most algorithms implemented under the Bayesian inference framework have no automatic modelfinder. In order to define a general model of DNA evolution, the rates of change need to be specified. The *clock model* defines how rates change globally across branches, and the additional *substitution model* determines the change of rates between characters. The clock model can be defined in three ways: *strict*, *relaxed*, or *no clock*. According to Dunn (2015b, p. 199), the strict clock model is inappropriate for linguistics, since the rate of change is constant across the tree. Defining a model with no clock indicates limitless evolutionary variation between branches, which is less likely to reflect a real evolutionary process (Dunn, 2015b). Therefore, the relaxed clock is the one of choice for linguistic data. The relaxed clock allows "rates to vary across the tree, chosen from a probability distribution whose mean is determined by the rate of the parent branch" (Dunn, 2015b, p. 199). For a substitution model, the *gamma model* is expected to be a good

fit on linguistic data (Dunn, 2015b). In this model, a gamma distribution is used to sample the rates for each class. Using a gamma distribution is popular, since it has the useful property of having a single parameter that determines the shape of the distribution (Huson et al., 2010). For this reason, it is expected to have the best fit on linguistic data, since different items in a word list can have a different stability (Dunn, 2015b).
The program used for concept tree reconstruction is *MrBayes* (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). MrBayes contains all possible combinations of models needed for a Bayesian analysis using MCMC on linguistic data. The concept trees are reconstructed using a relaxed clock and the gamma model to specify the rates of change. MrBayes provides several outputs. The most important ones for further analyses are the tree samples, including all trees according to the posterior distribution, and the consensus trees. The support values in the consensus tree are provided by the posterior distribution, i.e. no further analysis like bootstrap is needed.

4.2.1 State-of-the-Art: Cognate Clustering

The classical comparative method does not only aim at identifying shared patterns via sequence comparison, but also at identifying etymologically related words and grouping them into cognate classes. The process of establishing sound correspondences and cognate classes is circular, i.e. cognates can only be identified according to sound correspondences, and sound correspondences occur only in cognates (Jäger and List, 2016). Within the computational comparative method, sequence alignments are used to cluster words into cognate classes. Cognate classes represent the etymological relationship between the languages within one concept, and are optimal features for the creation of a binary character matrix.
NELex does not provide expert cognate judgments for the words in the basic concept list. Therefore, the cognates are clustered automatically using the OnlinePMI program by Rama et al. (2017). The PMI-based alignment method, combined with the Infomap clustering algorithm, assigns cognate classes to the words concept-wise. An illustration example is given in table 4.1. The algorithm assigns an index to each lexical item in the concept representing the cognate class. The shared cognate class indicates the etymological relationship, i.e. the descent from a common ancestor. The cognate classes serve as features to generate a binary character matrix.

Table 4.6 represents the language sample from table 4.1 with the transcriptions of the lexical items using the ASJP encoding, the assigned cognate class(es), and their binary coding. The data matrix is filled with 1 for presence and 0 for absence of the

Language	Phono. string	cc	Cognate sets		
			m	b	f
English	mauntin	m	1	0	0
Icelandic	fEtl	f	0	0	1
Swedish	bErg, fyEl	b,f	0	1	1
Norwegian	bErg, fyEl	b,f	0	1	1
Dutch	bErx	b	0	1	0
German	bErk	b	0	1	0
Romanian	munte	m	1	0	0
Italian	monta5a	m	1	0	0
French	mo,mota5	m	1	0	0
Catalan	munta53	m	1	0	0
Portuguese	mot3, mota5a	m	1	0	0
Spanish	monta5a	m	1	0	0

Tab. 4.6.: Binary (presence-absence) coding of cognate classes for the concept 'mountain'. The abbreviations stand for: Germanic languages `b` = 'berg', `f` = 'fjäll'; Romance languages `m` = 'mountain'.

cognate class. The integration of synonyms in a binary system is straightforward. In cases like Swedish and Norwegian, where the synonyms belong to different cognate classes, both cognate classes receive a 1, indicating their presence. In comparison, French and Portuguese contain two representations for the concept 'mountain' belonging to the same cognate class, which is indicated by assigning a 1 to the cognate class `m` indicating the presence. In this case, the synonyms cannot be captured in the representation of the binary matrix.

Language	Cognate sets		
	m	b	f
English	1	0	0
Icelandic	0	0	1
Swedish	0	1	1
Norwegian	0	1	1
Dutch	0	1	0
German	0	1	0
Romanian	1	0	0
Italian	1	0	0
French	1	0	0
Catalan	1	0	0
Portuguese	1	0	0
Spanish	1	0	0

Tab. 4.7.: Binary (presence-absence) character matrix for the concept 'mountain'.

In the final presence-absence matrix, displayed in table 4.7, each row represents a language and each column a cognate class, where 1 indicates presence and 0 absence, respectively. The number of columns depends on the number of cognate classes assigned by the clustering algorithm. For each concept, a binary presence-absence matrix is generated which serves as input to IQ-Tree and MrBayes for tree reconstruction.
Admittedly, the difficulty of binary matrix generation using cognate classes, i.e. the assignment of few cognate classes by the clustering algorithm, is still present in the analysis. The impact on the reconstruction of concept trees using a small character matrix with a low number of cognate classes is unclear. It is therefore advisable to include the state-of-the-art method for binary matrix generation in the analyses in order to get an impression on the effect of small character matrices in terms of concept tree reconstruction.

4.2.2 Multiple Sequence Alignment for Character Matrices

In terms of linguistic tree reconstruction, character-based methods allow a detailed and fine-grained inference on language classifications and change from pre-processed data. Distance-based methods, on the other hand, can be applied to raw, unprocessed data, resulting in equally good language classifications compared to an expert classification tree. The fundamental idea of the method proposed by Jäger and List (2015) is to combine the best of both methods, i.e. the processing of raw data with the detailed, fine-grained inference of linguistic trees. The approach uses *multiple sequence alignment* (MSA) on the raw and unprocessed data, out of which a binary character matrix is constructed for the fine-grained tree inference.
Multiple sequence alignment is applied to linguistic data to get an impression on the diversity of the words and additional evidence on sound correspondences from the complete sequence comparison. This is in accordance with the classical comparative method, where sound correspondences and cognates are identified via comparisons over multiple words in the languages. Within a MSA, all sequences present in the data are aligned in one block. Overall structures, similarities, and differences between all sequences under comparison can be seen at one sight. A multiple sequence alignment between all lexical items in one concept could indicate the direct relationships between the languages and their words. In example 4.2, the multiple alignments of the concept 'mountain' are illustrated for three Germanic languages and two Romance languages, plus English. The variety of sounds in the lexical items determines the amount of gaps needed for the alignment. Within the Germanic language sample the diversity of sounds is low, resulting in a small

amount of gaps needed to align the words. Within the alignment of the two Romance languages and English, the high diversity of sounds determines the higher number of gaps included in the MSA.
The major problem of multiple alignments is the increasing complexity of the analysis, which grows exponentially with the number of sequences being aligned (Jäger and List, 2016). Therefore, most algorithms use heuristic methods, such as *progressive alignment*, rather than global optimization to identify the optimal alignment between sequences.[19]

(4.2) Multiple alignments for the concept 'mountain'

a.
German:	b	E	r	k
Dutch:	b	E	r	x
Swedish:	b	E	r	g
...				

b.
English:	m	a	-	u	n	t	i	-	n	-
French:	m	-	o	-	-	t	-	a	5	-
Spanish:	m	-	o	-	n	t	-	a	5	a
...										

Progressive alignment requires a *guide tree* that captures the sequences in question and, ideally, the evolutionary relationships between them. The sequences are present at the leaves of the tree. The tree is traversed bottom-up and aligns the sequences of the daughter nodes at an inner node. The complete multiple sequence alignment is present at the root of the tree. The disadvantage of progressive alignment is that each node can only use the information which is present in the corresponding daughter nodes. If the alignment at the inner node is not optimal or etymologically correct, the decision cannot be undone further up in the tree. The *T-Coffee* algorithm (tree-based consistency objective function for alignment evaluation), introduced by Notredame et al. (2000), contains a pre-processing approach called *consistency-based alignments*, which is used to enhance the progressive alignments. The basic idea is that a good multiple alignment of the sequences should be consistent with a set of independently established pairwise alignments. Jäger and List (2015) combined the PMI-based method introduced for pairwise alignments with the T-Coffee algorithm, resulting in a two-stage PMI-T-Coffee approach. In the first step, a set of independently computed pairwise alignments, called a *library*, is generated using the

[19] The strategy of progressive alignment was introduced by Hogeweg and Hesper (1984) and Feng and Doolittle (1987).

PMI-based NW algorithm. Within the library, a scoring function is initialized, where each pairwise alignment receives a weight according to the matches and mismatches in it. In the second step, an extended library of composite alignments is collected, i.e. each pair of sequences is aligned with the rest of all sequences. The scoring function is further modified and used with the progressive alignment in order to compute the MSA at the inner nodes and the root of the tree.[20] The T-Coffee algorithm enhances the progressive alignment according to these pre-processing steps, and can therefore be used to avoid the problem that each node can only use the information of its daughters.

Although the field of bioinformatics provides several computational algorithms and approaches to align multiple sequences, the applications of MSA in linguistics is still in its infancy. The first algorithm on MSA of phonetic sequences was proposed by Covington (1998). However, the method contained an inefficient tree-search, which was not tested on large datasets (Jäger and List, 2016). Prokić et al. (2009) applied the *ALPHAMALIG algorithm* (Alonso et al., 2004) for cognate alignments in Bulgarian dialects to study discourse structure, which reached a high accuracy in the comparison to gold standard data. The two latest studies on MSA in linguistics are the one by List (2012) and Jäger and List (2015). List (2012) presented a progressive alignment implementation of the *SCA* (sound-class-based phonetic alignment) method (List, 2010), which outperforms the algorithm introduced by Prokić et al. (2009). Jäger and List (2015) showed that phylogenetic trees reconstructed with the PMI-T-Coffee algorithm came closer to the gold standard Glottolog tree than the SCA algorithm.

To construct binary character matrices, the lexical items for each concept in the NELex word list are aligned according to the PMI-T-Coffee algorithm of Jäger and List (2015). The guide tree needed for the progressive alignment is a language tree computed using the PMI-based NW algorithm and the aggregation of the language distances proposed by Jäger (2013b). Synonyms in the dataset can easily be taken into account, i.e. the MSA is constructed to align more than two words, and synonyms can simply be added to the alignment. The result is a complete multiple alignment of the lexical items for the corresponding concept. To construct a binary presence-absence matrix, the languages need to be categorized according to features, i.e. the MSA needs to be organized in a binary data format. Since the alignments are arranged in columns, as illustrated in example 4.2, binary characters can be created using the alignment information. Each occurrence of a sound class present in each column of the alignment block is transformed into a binary character for the languages, where 1 means presence and 0 absence of the character. Those binary characters are arranged in a matrix, where the languages are present in the rows

[20] A workflow of the algorithm is illustrated in Jäger and List (2015).

and the binary characters in the columns. For each concept, a binary data matrix is generated, and trees are reconstructed using IQ-Tree and MrBayes.

4.2.3 String Subsequences and Matrix Generation

Binary data matrices cannot only be reconstructed using a binary version of a MSA or cognate data, but also by using string subsequences. Phonetic strings or lexical items can be split into substrings of any length. The following approach was inspired by the work of Rama and Borin (2014), who used phonetic n-grams to establish phonological similarity with item stability. Earlier works, like the ones from Dunning (1994), Huffman (1998), and Singh and Surana (2007), motivated the use of n-grams for language identification tasks.
In computational linguistics, an n-gram is the result of splitting a sequence into n items. The n-gram is therefore a contiguous sequence of n characters from the transcription of the lexical items. Using Latin numerical prefixes, an n-gram of size one is called *unigram*, size two is a *bigram*, size three is a *trigram*, and so forth. Using NELex word lists, n-grams are generated by using phonological transcriptions of the lexical items, i.e. a substring is a sequence of phonological characters. For each language pair in a concept, the intersection of the substrings is built, out of which a binary data matrix is then reconstructed.
In the following, three different methods are introduced to generate n-gram sequences:

1. String subsequences using the traditional n-gram approach;
2. String subsequences using pairwise alignment;
3. String subsequences using pairwise alignment and n-grams.

The N-gram Approach In the first approach, n-grams of different sizes are created by splitting the phonological transcriptions of the words into n items. The resulting n-grams are unigrams, bigrams, and gappy bigrams. In gappy bigrams, one character is skipped during the automatic generation of bigrams. The size of n is kept small to avoid the issue of small character matrices. Additionally, some lexical items in the word list are not long enough to construct n-grams with a higher value for n. The gappy bigrams increase the amount of substrings in the n-gram set, which adds more information to the language comparison than raising the value of n. For bigrams and gappy bigrams, symbols to indicate the beginning (^) and ending ($) of a lexical item are added (Bergsma and Kondrak, 2007). Each lexical item in the concept is

split into n-grams to construct a set of all unigrams, bigrams, and gappy bigrams:

German bErk $\Rightarrow$ {b, E, r, k, ^b, bE, Er, rk, k$, ^E, br, Ek, r$}
Dutch bErx $\Rightarrow$ {b, E, r, x, ^b, bE, Er, rx, x$, ^E, br, Ex, r$}

The two sets of n-grams for the German and Dutch words for 'mountain' contain all unigrams, bigrams, and gappy bigrams. The intersection of the two n-gram sets indicates the set of matching substrings:

Matching substrings $\Rightarrow$ {b, E, r, ^b, bE, Er,^E, br, r$}

This introduces a similarity measure between the language pair in the corresponding concept, which indicates the relationship between the languages according to their shared substrings. Both German and Dutch have an n-gram set of size 13, where 9 n-grams are shared by both languages. In the case of synonyms, a shared n-gram set is built for each word pair. The shared substring of both languages is the intersection of all sets of possible word pairs. The data matrix therefore contains the languages in the rows and all possible n-grams in the columns. If an n-gram is a match between two substrings, the presence is coded with 1 and the absence is coded with 0. The character matrix for the set of shared n-grams between German and Dutch is displayed in table 4.8.

Language	b	E	r	k	x	^b	bE	Er	rk	rx	k$	x$	^E	br	Ek	Ex	r$
German	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	1
Dutch	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	1	1

Tab. 4.8.: Binary data matrix of n-gram subsequences between German and Dutch.

Pairwise Alignment The second approach uses the resulting string alignment of the PMI-based NW algorithm introduced in chapter 4.1.1 to construct binary character matrices. In contrast to the distance-based methods, where the automatic pairwise alignments are used to compute a linguistic distance between languages, this approach uses the actual string alignment to generate a set of substrings for the language comparison.

(4.3) Pairwise alignments of the words for the concept ‘mountain’

a.
German:	b	E	r	k
Dutch:	b	E	r	x

b.
German:	b	-	E	r	k	-	-
Icelandic:	-	f	E	-	-	t	l

c.
German:	b	-	E	r	x	-	-
Icelandic:	-	f	E	-	-	t	l

Considering all possible combinations of pairwise alignments between the three languages, a set of shared alignments indicates the similarity between them. For German and Dutch, the similarity is high, and no gaps need to be inserted in the alignment. Within the comparison of the two languages German and Dutch with the Icelandic word `fEtl`, the distance between the languages increases, which is indicated by the amount of gaps inserted in the alignment. The string alignments are used to build sets containing all aligned sounds, i.e. the columns of the alignment. The alignment between a sound and a gap is also taken into account. In the case of synonyms, all possible combinations of word pairs are aligned. Using the alignments between the language pairs, a binary data matrix is reconstructed, where 1 indicates presence and 0 absence of the sound correspondences. An example of a binary matrix using the three languages from the example is given in table 4.9.

Language	bb	EE	rr	kx	b-	-f	r-	k-	x-	-t	-l
German	1	1	1	1	1	0	1	1	0	1	1
Dutch	1	1	1	1	0	1	1	0	1	1	1
Icelandic	0	1	0	0	1	1	1	1	1	1	1

Tab. 4.9.: Binary data matrix of aligned strings between German, Dutch, and Icelandic.

Pairwise Alignment and N-grams The third approach is a combination of the two introduced approaches, i.e. alignment of the strings using the PMI-based NW algorithm, and splitting of the strings to generate n-grams.[21] In the first step, the lexical items are aligned using the PMI-based NW algorithm as illustrated in example (4.3). The strings from the alignment are used to create n-grams. In the alignment between German and Icelandic, gaps need to be inserted to align the phonetic characters. The string for the n-gram analysis would change for German from `bErk` into `bErk--` and for Icelandic from `fEtl` into `f-E--tl`. For the German-Icelandic pair, the n-gram

[21] Jäger and Wichmann (2016) combined the PMI distance with a bigram inventory distance to infer a world tree. The bigram inventory distance should add information where the lexical distance does not provide a detectable signal.

substrings would be as follows:

German `bErk--` ⇒
`{b, E, r, k, -, ^b, bE, Er, rk, k-, --, -$, ^E, br, Ek, r-,k- -$}`
Icelandic `f-E--tl` ⇒
`{f, E, -,t, l, ^f, f-, -E,E-,--, -t, tl, l$, ^-,fE,E-,-t,-l, t$}`

If a language contains synonyms, all possible word pairs are aligned, and a shared n-gram set is built. The intersection of the subsequences is used to reconstruct a binary matrix. The presence-absence matrix is similar to the one displayed in table 4.8, where 1 indicates presence and 0 absence of the n-grams.

All binary presence-absence matrices serve as input for IQ-Tree to reconstruct concept trees within the maximum likelihood framework and for MrBayes to reconstruct Bayesian concept trees.

4.2.4 Evaluation

The approaches to generate binary character matrices for concept tree reconstruction result in more than one matrix generation method. As already introduced for the distance-based methods, the aim is to determine the method with the most representative trees. In terms of character-based methods, the trees reconstructed using IQ-Tree are separately evaluated from the Bayesian trees obtained from MrBayes, i.e. resulting in one evaluation for each framework.
There are five different approaches to compile character matrices for the tree reconstructions with IQ-Tree and MrBayes, resulting in five matrix computations for each framework. All options are listed in table 4.10.

The first approach is state-of-the-art cognate clustering via the automatic clustering algorithm OnlinePMI, where a binary presence-absence matrix is compiled out of the obtained cognate classes. The second approach uses the PMI-T-Coffee algorithm to align multiple sequences. The columns of the resulting MSA are represented in a binary character matrix. The last three approaches use the idea of splitting strings into subsequences for the generation of presence-absence matrices. The subsequences can be obtained in three different ways: through n-gram splitting, decomposing a pairwise alignment, or n-gram computation of the pairwise alignment strings.

Method	Description
CT_{ML+cog}	IQ-Tree + cognate data
CT_{ML+msa}	IQ-Tree + PMI-T-Coffee
$CT_{ML+ngrams}$	IQ-Tree + subsequences with n-grams
CT_{ML+NW}	IQ-Tree + subsequences with PMI-based NW
$CT_{ML+ngramsNW}$	IQ-Tree + subsequences with PMI-based NW and n-grams
$CT_{Bayesian+cog}$	MrBayes + cognate data
$CT_{Bayesian+msa}$	MrBayes + PMI-T-Coffee
$CT_{Bayesian+ngrams}$	MrBayes + subsequences with n-grams
$CT_{Bayesian+NW}$	MrBayes + subsequences with PMI-based NW
$CT_{Bayesian+ngramsNW}$	MrBayes + subsequences with PMI-based NW and n-grams

Tab. 4.10.: The five approaches to reconstruct data matrices used for tree reconstructions with IQ-Tree and MrBayes.

As described for the distance-based methods in chapter 4.1.5, the maximum clade credibility method is used to determine the best tree representative in a tree sample and its clade credibility score. For each concept, the clade credibility scores are compared between all methods to identify the method with the highest score. The method with the highest sum of highest clade credibility scores for all concepts serves as the ideal method for further analyses. Since the evaluation is split into two parts, this analysis is done separately for each framework.

Evaluation Maximum Likelihood Within the maximum likelihood framework, the bootstrap replicates reconstructed with the implemented bootstrap technique in IQ-Tree serve as tree sample. For each method, the MCC tree with the highest clade credibility score is determined in the set of bootstrap replicates computed for each concept. Due to runtime issues, restrictions need to be made for the standard bootstrap method. For this evaluation task, the bootstrap sample for all methods is computed out of the most stable 200 concepts (Dellert and Buch, 2018) from NELex to create 100 bootstrap replicates.[22] The sums of the highest clade credibility scores for each method are listed in table 4.11.

[22] Nguyen et al. (2014) advertise the Ultrafast Bootstrap Method, introduced by Hoang et al. (2017), implemented in IQ-Tree. However, the method reduces the number of characters in a sample, which can result in a reduced number of leaves in the tree. The fact that the trees in the sample do not have the same set of leaves has a great impact on the maximum clade credibility method and the computation of the MCC tree. Therefore, it is more suitable to use the standard bootstrap method with restrictions for this analysis.

Method	SCC
CT_{ML+cog}	30
CT_{ML+msa}	3
$CT_{ML+ngrams}$	47
CT_{ML+NW}	50
$CT_{ML+ngramsNW}$	**70**

Tab. 4.11.: The MCC comparison of the five approaches to reconstruct data matrices used with IQ-Tree. The most optimal method is marked in bold.

Evaluation Bayesian Inference For the Bayesian analysis with MrBayes, the tree sample for one concept consists out of the trees contained in the posterior distribution. The number of posterior trees can vary between concepts. However, this does not have an impact on the evaluation of the methods using the MCC method. For the MCC analysis, the first 25% of the posterior trees from the burn-in are discarded. The rest of trees are used to determine the MCC tree with the highest score for each concept in each method. The method with the highest sum of clade credibility scores serves as the ideal method. The results are listed in table 4.12.

Method	SCC
$CT_{Bayesian+cog}$	0
$CT_{Bayesian+msa}$	0
$CT_{Bayesian+ngrams}$	1
$CT_{Bayesian+NW}$	**1013**
$CT_{Bayesian+ngramsNW}$	2

Tab. 4.12.: The MCC comparison of the five approaches to reconstruct data matrices used with MrBayes. The most optimal method is marked in bold.

Discussion In both analyses, the approaches using automatically inferred cognate classes and multiple alignment via PMI-T-Coffee are outperformed by the other methods using subsequences. The results for the tree reconstructions using automatic cognate clustering underline the assumption that the reconstruction of concept trees depends on the dimension of the character matrix. Although Rama et al. (2018) showed that language tree reconstruction in a Bayesian framework using automatically inferred cognate classes results in reliable trees, this effect is not mirrored in the reconstruction of concept trees. It cannot be excluded that the OnlinePMI algorithm assigns few cognate classes to the words within one concept. The basic vocabulary list contains stable meanings indicating the relatedness of the words

between the languages, where most of the words are inherited. This results in the clustering of words into a small set of cognate classes, which is reflected in the character matrix. The effect of small character matrices on character-based tree reconstruction is unclear. However, the evaluation results show a clear tendency towards the other methods under comparison. Multiple sequence alignment using PMI-T-Coffee on a concept level does not mirror the reliable results shown for phylogenetic reconstruction on language data by Jäger and List (2015). The automatically inferred language tree used as guide tree can lead to spurious results in the bottom-up alignment, since the words of a concept are placed directly at the leaves and are aligned with their sister language in the language tree. In the reconstruction of language trees, this is an advantage over the pairwise sequence alignment. In the reconstruction of concept trees, pairwise sequence alignment represented in a binary way leads to better results. However, using a language tree is obviously the best choice, since there are no expert concept trees and, up to this point, there is no assessment of automatically inferred concept trees. In the Bayesian analysis, the subsequences received from the PMI-weighted NW algorithm have by far the best results according to the MCC analysis. In the maximum likelihood framework, the subsequences preserved from the method using the PMI-based NW algorithm in combination with the n-gram analysis outperforms the other methods. The underlying idea of using subsequences for a binary presence-absence coding is to receive more fine-grained information from the data and extend the columns in the matrix for a better character-based tree reconstruction. In comparison to the cognate clustering and MSA approach, using subsequences works well on linguistic data for the reconstruction of concept trees.

5 Networks: Modeling Horizontal Word Transfer

The network model is a representation combining the tree and the wave model in order to shed light on a vast majority of the evolution of languages, including inheritance and horizontal transmission. Evolutionary networks are the networks of choice to illustrate language evolution and horizontal transfer according to evolutionary processes. Explicitly, HGT networks can capture evolutionary events due to borrowing, including information on the loaned word, the RL, the SL, and the direction of transfer. During the process of horizontal word transfer, only the recipient language is modified, which affects the evolutionary history of the word. This is mirrored in the concept trees, which are reconciled with a language tree in order to detect the discordance between the two trees. The mismatches in the tree determine the number of reticulations and therefore also the detected transfer events. Phylogenetic tree reconciliation algorithms can only detect potential HWT events, which can result in arbitrary reticulations in the network. Since less attention is paid to the parallels between HGT and borrowing in CHL, it is worthwhile to adapt a phylogenetic HGT algorithm to linguistic data to gain a detailed estimation of the application, performance, and results obtained from the algorithm. The following analyses should indicate whether the direct transfer of HGT algorithms into CHL is fruitful for the detection of borrowings and the corresponding loanwords.

The algorithm of choice from phylogenetics is the one developed by Boc and Makarenkov (2003), which was adapted into CHL to detect loanwords and model HWT networks. The decision of choosing this phylogenetic algorithm over others is explained in the following chapter along with the underlying computational model. The suitability of the adaptation and application of the HGT method is examined in the evaluation compared to other linguistic approaches.

In linguistics, Delz (2014) introduced a tree comparison method based on the idea of tree reconciliation to detect loanwords, which can be applied to the data from NELex to serve as direct comparison to the HGT algorithm. The MLN approach, originally introduced in phylogenetics by Dagan and Martin (2007), was applied to the task of borrowing detection by Nelson-Sathi et al. (2011). To obtain transfer scenarios of lexical items, the method can be applied to concept data to detect the lexical items included in the contact scenario. The third linguistic model is the PLFI

model developed by Dellert (2019a). The detected contact situations by the PLFI method contain information on specific borrowing events and lexical items involved in the transfer, which can be statistically evaluated.
First, the HGT algorithm by (Makarenkov et al., 2006) and the additional three linguistics-based algorithms are introduced in more detail. Both the underlying concepts and the advantages and disadvantages are presented in terms of loanword detection. Second, the preconditions for the tree reconciliation algorithms, the MLN, and the PLFI are introduced. For all algorithms, equal pre-processing steps are necessary to maintain an appropriate evaluation.

5.1 Algorithms for Horizontal Transfer

The evolutionary process of borrowing consists of several phases (i.e. language contact, speaker interaction, borrowing, adaptation, and integration), where each phase contains different tasks. To get a complete picture of the whole borrowing process, a deeper understanding of the different stages is needed to explain the transfer and its results. The complexity of the process leads to various challenges when it comes to loanword identification. Not only is their adaptation and integration process dependent on the single transfer, but also on the detection of the donor language. Less, if any, universal assumptions can be made to account for various transfer processes between a language pair, not to mention the difficulty of developing universal assumptions between several language pairs.
From an algorithmic perspective, the computational comparative method can function as an underlying method in order to use cognates and word alignments as criteria to distinguish an inherited word from a loanword. This is the key assumption for the HWT algorithms introduced below. The visualization of horizontal transfer within a network model combines the language classifications obtained from the computational comparative method with transfer scenarios. The main aim of the algorithms is to detect horizontal word transfer events, including the loaned word, the RL, the SL, and the direction of the transfer. Each algorithm has advantages and disadvantages over the others, and only two out of four can model the direction of the transfer.
From a linguistic point of view, the process of borrowing is complex and contains additional parameters for a complete reconstruction of the borrowing scenario. However, the construction and implementation of such a full-featured borrowing model taking all parameters into account is not realizable up to this point. More analyses need to be made in both linguistics and CHL to get a better understanding

of modeling language contact and lexical transfer. The following algorithms serve as first steps into this direction to gain insights in the application and performance of different computational approaches to model language contact and lexical borrowing. The results provide added value to further work on loanword detection and modeling transfer scenarios.

5.1.1 Phylogenetic HWT Algorithm

The HGT algorithm was developed by Makarenkov et al. (2006) and extended by Boc et al. (2010b) to detect horizontal gene transfer events in the evolution of microorganisms.[1] Due to the parallels between horizontal gene transfer and borrowing, the algorithm is adapted and applied to linguistic data to detect loanwords. The tree reconciliation procedure, introduced in chapter 3.3, serves as an underlying idea, where a concept tree is mapped into a language tree to estimate the discordance between the two trees. The main advantage of the algorithm is the additional usage of tree replicates to assign a reliability value to the horizontal transfer events. Delz (2013) discussed the advantages of the application of the HGT algorithm to linguistic data. The results obtained by the method include the loaned word, the RL, the SL, and the direction of transfer. In this preliminary study, the main aim was to elaborate the proposed parallel between HGT and borrowing. The thesis was underlined by a small experiment on using the algorithm to detect horizontal word transfer events. Since the reconstruction of concept trees was a secondary effect of the study, the algorithm was not tested on different reconstructed concept trees along with tree replicates to determine transfer events with reliability values. By using the innovatively computed concept trees presented in chapter 4 along with tree replicates, the reliability of the transfer events can be estimated.

Description of the Algorithm

Boc et al. (2010b) described a new polynomial-time algorithm to infer HGT events, which is adapted to CHL to detect borrowings. The algorithm searches for an optimal scenario of *subtree prune and regraft* (SPR) moves to transform a concept tree into a language tree. In addition, four optimization criteria are introduced for the selection of the optimal SPR move, i.e. HGT event. After introducing SPR and the four

[1] The phylogenetic algorithm is integrated in the online platform *Trex-online* (Boc et al., 2012), which is accessible under `http://trex.uqam.ca/`.

optimization criteria, the preliminary and computational steps of the algorithm are explained.[2]

SPR Subtree prune and regraft (SPR) is a branch-swapping method where a subtree from a given phylogenetic tree is detached ("pruned") and reattached ("regrafted") at a different position in the tree (Huson et al., 2010).[3] One operation of pruning and regrafting a subtree is called SPR move. A branch-swapping method like SPR can also be used to define a distance between two phylogenetic trees, i.e. the minimum number of SPR moves needed to transform one tree into the other (Huson et al., 2010). The SPR distance determines the discordance between the concept and the language tree. The language tree is gradually transformed into the concept tree by a series of SPR moves. The result of each SPR move is a transformed tree, which serves as new tree for the comparison. The goal is to find the minimum number of SPR moves, i.e. the shortest sequence of transformed trees needed to transform the language tree into the concept tree. This determines the SPR distance between the two trees, which is equal to the number of transfer events detected by the algorithm. However, at each transformation step, several SPR moves are possible, i.e. there is more than one possibility to reattach the subtree to the language tree. The different possibilities of reattachment are evaluated using one of four optimization criteria in order to find the optimal SPR move.

Optimization Criteria The first criterion is the *least-square* (LS) function, which is calculated using the patristic distance between two leaves in both the language and the concept tree.[4] The second criterion is the topological *Robinson-Foulds* (RF) distance (Robinson and Foulds, 1981), which is based on the symmetric differences of the bipartitions in the tree. It computes the minimum number of elementary operations (e.g. merging and splitting of nodes) to transform one tree into the other one. The third criterion is the quartet distance (QD) (Estabrook et al., 1985), introduced in chapter 4.1.4, which is determined by the number of quartets that differ between the two compared trees. The fourth criterion is the *bipartition dissimilarity* (BD) between two phylogenies, which is introduced in Boc et al. (2010b) along with the HGT algorithm. A presupposition of computing the BD between two trees is that both trees are binary branching and have the same set of leaves. In figure 5.1, two unrooted sample trees and their corresponding bipartition tables are illustrated. Each row in the bipartition table indicates a bipartition vector,

[2] See Boc et al. (2010b) for a detailed explanation including all mathematical formula.
[3] Huson et al. (2010, p. 39) give a graphical illustration of the SPR operation.
[4] A mathematical description and the formula is given in Boc et al. (2010b, p. 196).

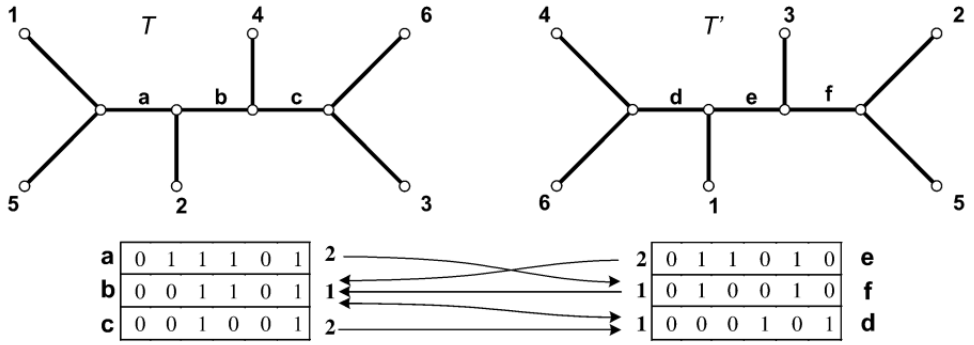


Fig. 5.1.: Two trees and their bipartion tables, taken from Boc et al. (2010b). Each row in the table corresponds to one internal branch of the tree. The arrows indicate the associations between the vectors. The values display the associated distance.

i.e. a binary vector induced by one internal branch of the tree. The lower-case letters next to the vector in figure 5.1 correspond to the lower-case letters in the tree, indicating a split to form a bipartition. The binary values 0 and 1 represent the leaves present in each partition. For example, the tree **T** is split at the internal branch **a**. The set of leaves $\{1, 5\}$ included in one partition is represented using 0. The other set of leaves $\{2, 3, 4, 6\}$ is represented by 1. For each tree, one bipartition table is created to demonstrate the possible splits at the internal branches of the tree. The bipartition dissimilarity between two tables, i.e. two trees, is computed using the formula given in Boc et al. (2010b, p. 197). The underlying idea is to compare each row of one table to all rows of the other using the *Hamming distance*. The Hamming distance compares two strings of equal length and determines the number of positions at which the symbols differ. The formula introduced in Boc et al. (2010b) also considers the complement vectors, where the values are switched to account for differences in the assignments of 0 and 1. In figure 5.1, each row of table one is compared to each row and its complement in table two, taking the minimal value as result, and vice versa. The numbers next to the rows in the figure 5.1 represent the result of the bipartition dissimilarity computed for the corresponding vector. The two vectors with the smallest distance are indicated by an arrow, as shown in figure 5.1. For example, vector **a** is compared to all vectors and their complements in table two, resulting in a bipartition dissimilarity of 2. The smallest Hamming distance is computed between vector **a** and **f**, which is indicated by an arrow. The overall bipartition dissimilarity of two trees is computed by summing over all minimum values obtained from the comparison and normalizing it by the number of tree typologies. In figure 5.1, the overall bipartition dissimilarity is $((2 + 1 + 2) + (2 + 1 + 1))/2 = 4.5$.

Optimization Criteria as Selection Parameter The optimization criteria serve as selection parameter to determine the optimal SPR move in the set of possible reattachment options. Each possible reattachment is performed to create a transformed tree. The transformed language tree is compared to the concept tree using one of the proposed criteria to compute a distance between the trees. The SPR move, which provides the minimum value of the selected criterion computed for the transformed language tree and the concept tree, is retained (Boc et al., 2010b).
In a simulation study, Boc et al. (2010b) examine the performance of the HGT algorithm according to the four optimization criteria and the resulting number of HGTs. The BD criterion obtained generally more accurate results than the other three optimization criteria, i.e. the BD-based strategy outperforms the other criteria in most simulation analyses.

Preliminary Steps The preliminary steps of the algorithm are summarized as follows:

1. Inference of the language and concept tree, denoted T and T' respectively, whose leaves are labeled with the same set of languages.
2. Rooting of the trees.
3. Reducing the size of the problem if there are identical subtrees in T and T' with two or more leaves. (This is done by the algorithm.)

The inference of concept trees was discussed in chapter 4. The three resulting optimal methods evaluated for distance and character-based tree reconstructions serve as input for the HGT algorithm. The reconstruction of the language tree is explained in chapter 5.2, along with the rooting of the trees. The tree rooting is essential in tree reconciliation. A misplaced root in the language or the concept tree can lead to false positive and false negative HWTs. If the trees are not rooted beforehand, the trees are rooted internally by the algorithm. Identical subtrees are recognized by the algorithm and excluded from the analysis, which reduces the problem of tree reconciliation. If all preliminary steps are carried out, the algorithm starts with the detection of HWT events.

HWT Detection The algorithm recursively performs the procedure described in the following to determine HWT events. Each step k equals one SPR move, including the different possibilities of reattachment on the tree.

1. Consider all possible HWTs between the pairs of the branches in the language tree T_{k-1}, where k is the number of steps, i.e. T_0 at step 1.

2. Reject all transfers between adjacent branches and those violating the subtree constraint.
3. Among them, find those satisfying theorem 2 first and theorem 1 second, and carry out the SPR moves.
4. Carry out all remaining SPR moves satisfying the subtree constraint, which transforms the tree from T_{k-1} to T_k.
5. The direction of each HWT is determined using the selected optimization criterion.
6. Reduce the problem by collapsing identical subtrees.

In the first step, all possible HWTs are considered, i.e. all possible SPR moves to transform the language tree. During this step, for each SPR move (i.e. each HWT) all regrafting possibilities are evaluated using the optimization criterion to find the optimal SPR move. At each step k, multiple SPR moves (i.e. multiple HWTs) can be carried out to transform the tree. Transfers between adjacent branches are rejected along with those violating the subtree constraint. The subtree constraint states that a transfer between branches is only allowed if, and only if, the newly emerged cluster rooted by that branch is already present in the concept tree. In other words, a fusion of two clades under one mother node by an SPR move is only possible if this newly formed clade is already present in the concept tree.[5] This constraint allows to arrange the topological conflicts between the two trees which are due to HWTs between close ancestors of the contemporary languages in the first step, and transfers that appeared deeper in the phylogeny in the second step. The two theorems defined in Boc et al. (2010b) specify some properties of the bipartitions in connection with HWTs satisfying the constraint. All SPR moves are carried out to transform the language tree T_{k-1} to T_k. For each HWT, the optimization criterion is used to identify the direction of the transfer. In the case of two opposite HWTs, the transfer that minimizes the value of the selected optimization criterion computed for the transformed language tree T_k and the concept tree T' is chosen. In the last step, the algorithm reduces the problem by collapsing the identical subtrees. This procedure is repeated recursively until the stopping criterion is met, i.e. when the coefficient of the optimization criterion equals zero. After the reconciliation of the language and concept tree, a backward procedure is carried out. This step eliminates redundant transfers, i.e. the removal of the transfer from the scenario, if it does not change the topology of the resulting concept tree. The remaining HWTs are the transfer of lexical items found by the algorithm for a specific concept, i.e. loanwords.

[5] See Boc et al. (2010b) for a visual representation.

Reliability of Transfer Events The main advantage of this algorithm is the usage of tree replicates to assign reliability values to the transfer events. In the traditional bootstrap analysis, introduced in chapter 4.1.4, confidence intervals are placed at the inner nodes of a consensus tree, indicating their stability. In the HGT algorithm, the bootstrap validation procedure, introduced in Makarenkov et al. (2006), is extended to assess reliability values of the determined HWTs. In an initial step, the concept tree replicates are inferred, and a tree sample is created. All HWT events detected in the original scenario are verified if they appear in the set of HWTs determined by using the replicated trees from the sample. This is done via a comparison of the corresponding SPR moves. In their study, Boc et al. (2010b) considered two HWT events as equal if the biparitions of both recipient and donor branch are equivalent in both transfers. This means the topologies of the recipient and donor subtrees can be different, even though the languages were the same in both compared HWTs. Boc et al. (2010b, p. 200) stated that "the bootstrap score of a [HWT] scenario can be defined as the product of all individual bootstrap scores found for the transfers being part of this scenario." The corresponding formula to compute the bootstrap score for a transfer event is given in Boc et al. (2010b, p. 200).

Evidence for the Choice of the Algorithm

There are several algorithms which can infer HWT events using the SPR distance. The most popular ones, next to the algorithm of Boc et al. (2010b), are *LatTrans*, *RIATA-HGT*, *HorizStory*, and *EEEP*. The LatTrans algorithm was introduced by Hallett and Lagergren (2001). The method can map numerous gene trees into a species tree in order to model HGTs. Although it can generate all shortest SPR scenarios, it is exponential in the number of HGT events. The HorizStory algorithm, described by MacLeod et al. (2005), intends to approximate the SPR distance between two rooted phylogenetic trees. In the first step, the algorithm rejects identical rooted subtrees present in both trees. Second, it carries out SPR moves on the remaining trees until they are reconciliated. Beiko and Hamilton (2006) introduced the efficient evaluation of edit paths (EEEP) algorithm, which determines the minimum number of SPR moves between two rooted trees. The key of this algorithm is the division of the species trees into those bipartitions that are concordant and discordant regarding the gene tree. The simulation study by Beiko and Hamilton (2006) to compare EEEP with HorizStory and LatTrans showed that LatTrans clearly outperforms the others in terms of HGT detection accuracy. The RIATA-HGT algorithm was developed by Nakhleh et al. (2005a) and is based on a divide-and-conquer approach. In simulation studies by Than et al. (2007), RIATA-HGT outperforms HorizStory, EEEP, and Lat-

Trans in terms of speed, while performing at least equally good in terms of accuracy. In the latest version of RIATA-HGT, Than and Nakhleh (2008) demonstrated that RIATA-HGT is faster and almost equivalent regarding accuracy compared to the LatTrans algorithm. However, LatTrans is still one of the most popular methods for HGT detection next to RIATA-HGT and the HGT algorithm of Boc et al. (2010b). Therefore, Boc et al. (2010b) carried out several simulation studies to compare the different algorithms.[6] The basis for the simulation studies are a species tree, a gene tree, and tree replicates. The tree replicates are reconstructed with respect to the number of HGT events in order to control for them in the evaluation of the comparison.

There are four different versions of the HGT algorithm due to the proposed optimization strategies by Boc et al. (2010b). In order to compare the best version of the HGT algorithm to LatTrans and RIATA-HGT, the optimization criterion with the best performance needs to be established. Hence, two simulation studies are carried out by Boc et al. (2010b) to compare the four optimization strategies. In the first simulation study, the optimization criteria are compared in terms of HGT detection rate (i.e. true positives), measured as a percentage, and number of HGT events. The BD-based algorithm yields very stable results, and mostly outperforms the other three optimization criteria. In the second study, the behavior of the four criteria is studied under the condition of correctness of the gene tree. The algorithm based on BD clearly outperforms the other three strategies. Therefore, Boc et al. (2010b) used the BD-based version of the algorithm for further comparisons with LatTrans and RIATA-HGT.

Boc et al. (2010b) presented a detailed comparison to the LatTrans algorithm regarding HGT detection accuracy and running time. In most of the studies, the BD-based HGT algorithm outperforms LatTrans in terms of HGT detection accuracy. The polynomial time complexity of the BD-based algorithm provides a significant gain in the running time compared to the exponential-time LatTrans algorithm. According to the improvement of the detection results in the simulation studies and the gain in running time, the BD-based HGT algorithm is preferred over LatTrans. In addition, the LatTrans algorithm does not provide the usage of tree replicates to establish reliability values of the transfer events.

The only algorithm to assess reliability values to transfer events is the RIATA-HGT algorithm introduced by Nakhleh et al. (2005a). The *PhyloNet* package (Than et al., 2008b) provides an extended implementation of RIATA-HGT, including the estimation of bootstrap support of HGT branches. The algorithms are compared in terms of HGT detection accuracy and running time. In terms of running time, the BD-based

[6] See Boc et al. (2010b) for a detailed description of the simulation studies, including the design, results, and graphical illustrations of the comparison.

HGT algorithm clearly outperforms RIATA-HGT, whereas for the HGT detection accuracy both algorithms provide equally good results. In addition, example studies using biological datasets are carried out using both algorithms. The results show that the BD-based HGT algorithm detects less and more accurate transfer events compared to RIATA-HGT. Than et al. (2008b) proposed a method to establish bootstrap support to internal HGT branches.[7] The support of a HGT branch is defined as "the maximum bootstrap support of all internal branches of the path linking the nodes [...] in the gene tree" (Boc et al., 2010b, p. 204), which can only be assigned if the gene trees, and ideally also the species tree, have bootstrap values pre-assigned to the trees. In comparison, the approach described in Boc et al. (2010b) uses tree replicates to assess the bootstrap values of HGT events, i.e. the method takes the different topologies of the replicated trees into account. Using tree replicates to compute the bootstrap values can include important information which might not be present in a unique gene tree with given bootstrap scores. As a result, the method proposed by Than et al. (2008b) can overestimate the bootstrap values, whereas the method by Boc et al. (2010b) uses more detailed information to determine realistic reliability scores. Boc et al. (2010b) evaluated the assignment and computation of the bootstrap values within the example studies using biological datasets, where the bootstrap scores found by RIATA-HGT are generally higher than the ones found by the BD-based algorithm.

In summary, the HGT events detected by the BD-based HGT algorithm improve the results in terms of accuracy and running time compared to the other algorithms. The estimation of bootstrap scores to verify the transfer events using tree replicates is a great advantage over the other algorithms. In addition, the resulting network contains the species tree as underlying tree and additional reticulate edges indicating the transfer events, including reliability values.

Linguistic Application of the BD-based HGT Algorithm

The adaptation of the BD-based HGT algorithm for linguistic tree reconciliation is introduced in chapter 3.3. The resulting network contains the language tree as an underlying representation of the language classifications and the reticulate edges to indicate the transfer events (i.e. SPR moves). The discordance between the trees is estimated using the SPR distance, i.e. the minimum number of SPR moves to transform the concept tree into the language tree. The evolutionary constraints, included in the BD-based algorithm, prohibit transfers between branches located at

[7] A description of the method to compute bootstrap support in RIATA-HGT is given in Than et al. (2008b) and in the supplementary material of Boc et al. (2010b).

the same lineage and crossing transfers, which lead to inappropriate HGT scenarios (Boc et al., 2010b).[8] The constraints ensure that the detected HGT events are in accordance with the characteristics of the HWT process, i.e. transfer events should only occur between contemporaneous languages and can affect both closely and distantly related languages. Borrowing modifies a single RL by adding new linguistic material. The transfer of a lexical item affects the evolutionary history of the word captured in the concept tree. The reconciliation of a concept tree to a language tree reveals this modification by computing the discordance between the trees, i.e. identifying the horizontal transmission. If the two trees are in concordance with each other, i.e. no horizontal transfer is present, the BD-based algorithm results in the language tree without reticulate edges. Otherwise, a network will be provided. In the following example, the language tree in figure 3.2 and the concept tree for 'mountain' in figure 3.3 are reconciled by the BD-based HGT algorithm. The resulting network is displayed in figure 5.2.

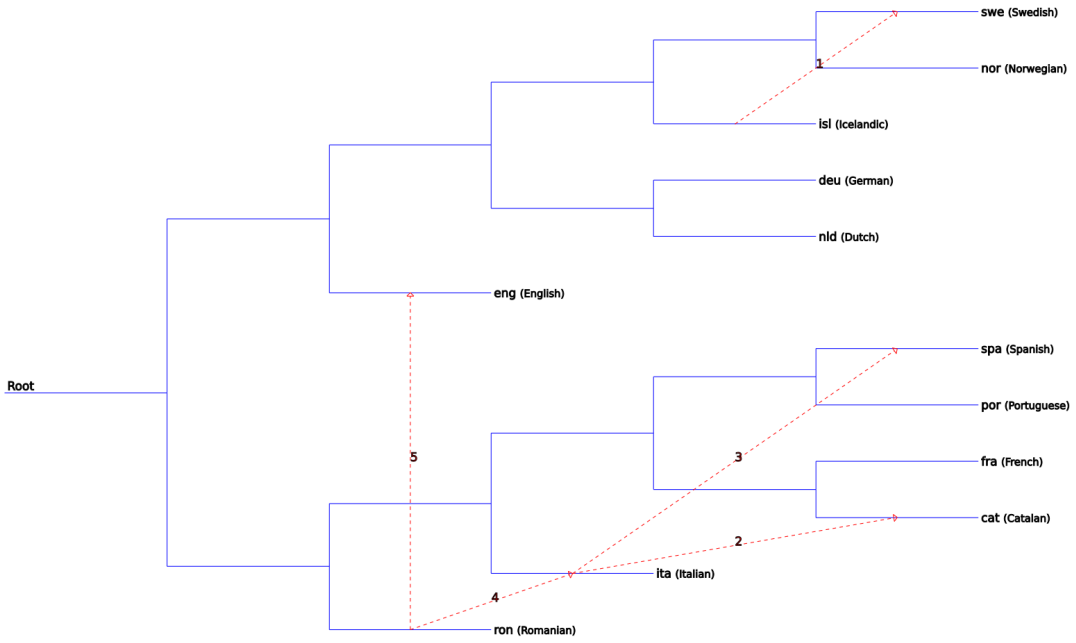


Fig. 5.2.: HGT network produced by the BD-based HGT algorithm for the selection of Germanic and Romance languages and the concept 'mountain'.

The network contains the language tree as an underlying graph, where reticulate edges are added to illustrate transfer events (i.e. SPR moves). The algorithm detects five transfer events, where one is between two Germanic languages (Swedish and Icelandic), three are within the Romance languages, and one is the transfer between a Romance language and English. It is not surprising that the algorithm identifies

[8] Descriptions of the evolutionary constraints are given in Boc et al. (2010b). See also Maddison (1997) as well as Page and Charleston (1998).

borrowings within a language family. There are known interborrowings within the Germanic languages as well as within the Romance languages (McMahon and McMahon, 2005). For the words of the concept 'mountain', no internal borrowing is recorded for either the Germanic or Romance transfers. On the one hand, it can be difficult to detect interfamily borrowing in a classical historical analysis. It might be the case that either language contact or language change are responsible for an approximation of the pronunciation of the words. Such a similarity can be reflected in the concept tree. On the other hand, the four interborrowings could be arbitrary events detected by the algorithm. Those can either be an artifact of the concept tree reconstruction, or due to the underlying heuristic search algorithm. The reliability values of the transfer events provided by the implemented bootstrap procedure can be used to reject arbitrary events, i.e. transfers with a low bootstrap support. This does not exclude the possibility that deficient concept trees cannot have an effect on the tree reconciliation. Quite the contrary is the case, since concept trees have a high impact on the reconciliation algorithm. However, the extent of deficient concept trees and their impact on the tree reconciliation algorithm is unclear. The evaluation of the complete data using the gold standard in chapter 6 is used to shed light on this issue.

Figure 5.3 displays the network including reliability values for the HGT events.[9] In the analysis, tree replicates from the concept tree are considered, while the language and concept tree were held constant. The support values indicate to what percentage the corresponding transfer can also be found in the reconciliation between the language tree and the concept tree replicates. The reliability values on the HGT events show that the support for internal borrowings within the language families is lower as the value for the borrowing between a Romance language and English. The bootstrap procedure can be used to reject borrowings with low support from the evaluation. However, for linguistic data, it is unclear up to which percentage the transfer events are seen as artifacts and when as reliable transfers. This is tested in the chapter 6.3 using different cut-off points to split the transfer events into two groups and evaluate the group with the higher percentages against the gold standard. The evaluation will shed light on the application of bootstrap values in order to reject arbitrary events from the analysis.

[9] Since the BD-based HGT algorithm uses heuristic methods to reconcile the two trees, the transfer events detected with and without assessment of reliability scores can differ.

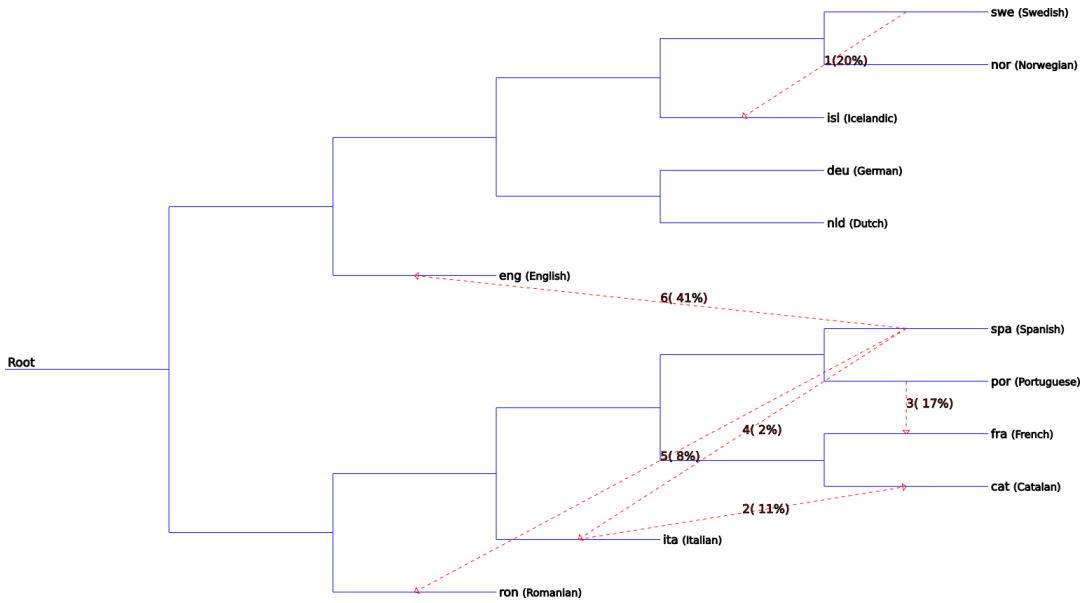


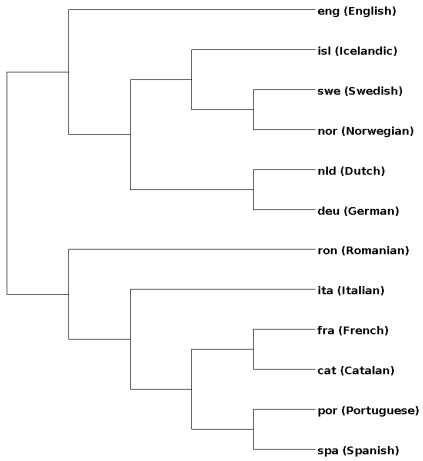
Fig. 5.3.: HGT network produced by the BD-based HGT algorithm containing reliability scores for the transfer events.

5.1.2 Loanword Detection through Tree Comparison

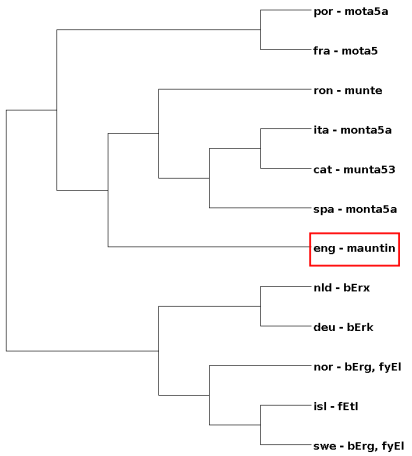
A linguistic tree comparison algorithm for loanword detection is developed by Delz (2014). The basic idea of the tree comparison method (TC) is to detect evolutionary events via computing the distance between a language and a concept tree. To see whether a language is affected by horizontal transmission, an approach inspired by the jackknife technique, introduced in chapter 4.1.4, is used. For detecting loanwords through tree comparisons, Delz (2014) uses the jackknife technique in order to generate subtrees of the sample to measure their distance. Given a language sample of size n, the generated subtrees are of size $n - 1$. The resulting sample consists of a language and a concept tree including all n languages, and n language and concept jackknife trees over $n - 1$ languages each. The trees in the sample are compared to each other in order to measure the discordance between them. Let d be the distance between the language and the concept tree for the complete sample and d' be the distance for a jackknife sample where language L is removed. If the criterion $d' < d$ holds, this means that the omission of L increases the fit of the trees. This is an indicator that an evolutionary event involving horizontal transfer has occurred for this concept in L. For each concept, the criterion provides a list of languages involved in borrowings, i.e. recipient languages.

Delz (2014) compared different distance measures for the comparison of the trees, i.e. the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981), the GQD introduced in chapter 4, and the *normalized triplet distance* (NTD) (Critchlow et al.,

1996). In correspondence to the results of a small study, Delz (2014) chose the triplet distance as optimal distance measure for the comparison of two trees according to the stability and accuracy of the results.[10] A triplet is the smallest informative subtree for binary rooted trees (Sand et al., 2013). The triplet distance is measured by splitting two trees into its triplets and counting the identical topologies induced by the three leaves in the two trees. For two rooted trees over the same set of leaves, the triplet distance is defined as the proportion of triplets over the leaves whose topology differs in the two trees. The triplet distance is normalized by the number of possible triplets over their leaves, which is estimated through the binomial, i.e. for n leaves, the number of possible triplets is $\binom{n}{3}$.
The following example shows the workflow of the TC algorithm using the language tree in figure 3.2 and the concept tree for 'mountain' in figure 3.3, repeated in figure 5.4. The pruned language and concept trees without English are displayed in figure 5.5.



(a) Language tree representing the classification of the Germanic and Romance languages.

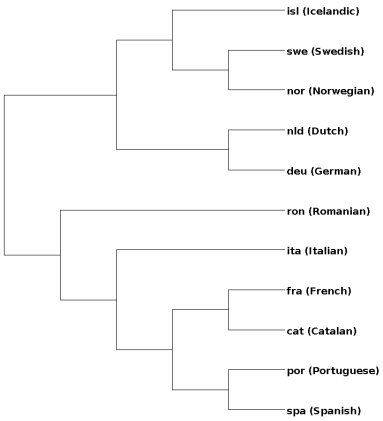


(b) Concept tree 'mountain' for the Germanic and Romance languages.

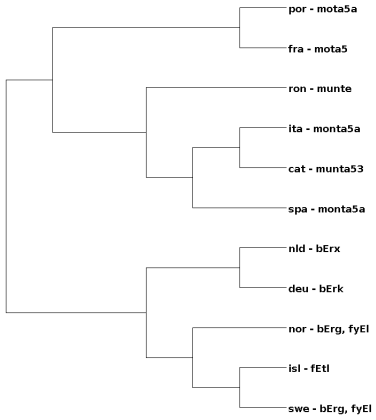
Fig. 5.4.: Illustration of language and a concept tree for a selection of Germanic and Romance languages.

In the first step, the NTD for the language and the concept tree, including the complete language sample, is computed. The result of $d = 0.2454$ shows the inconsistency between the two trees, which is due to the different clustering of the languages. In the second step, the NTD is computed between the pruned language tree and the jackknife concept tree, where English is removed. The resulting distance is $d' = 0.096$. Since the criterion $d' < d$ is true, it is assumed that the English word *mountain* is a loanword. This step is repeated for all languages in the sample. The

[10] See Delz (2014) for a detailed description of the distances and the example study.



(a) Language tree representing the classification of the languages in the sample without English.



(b) Concept tree 'mountain' for the language sample without English.

Fig. 5.5.: Illustration of a jackknife replicate language and a concept tree for a selection of Germanic and Romance languages without English.

result is a candidate list including all languages for which the criterion holds. A network is used to display the results, where the language tree is the underlying structure and reticulate edges indicate the connection to the sister node in the concept tree.[11] The TC algorithm does not include the detection of the SL and the direction of the transfer.

In an unpublished study, Koellner and Dellert (2017) combined the tree comparison study with different linguistics-based *ancestral state reconstruction* (ASR) techniques introduced in Koellner and Dellert (2016). The fundamental idea was to filter out arbitrary events which are not due to borrowing. However, on a small test and evaluation sample, the filters did not lead to the expected improvements of results. The basic tree comparison algorithm using the NTD still leads to the best results, which is why the other options are rejected in the following analyses.

The TC algorithm can be extended in order to compute reliability scores for the transfer events, i.e. the detected loanwords. The procedure is similar to the one implemented in Boc et al. (2010b), introduced in chapter 5.1.1. The tree comparison algorithm is applied to the language and the concept tree. In the bootstrap analysis, this procedure is repeated using the concept tree replicates, whereas the language tree is held constant. All languages detected as RLs in the original scenario are verified if they appear in the set of languages determined by using the replicated trees from the sample. The appearances of each language as a RL in the replicated trees are counted, and the result is normalized by the number of tree replicates used in the analysis. The resulting percentage score indicates the frequency of occurrence,

[11] See Delz (2014) for a graphical representation of the network.

i.e. the stability of appearance of the transfer event. The reliability score can be used to filter out arbitrary events, i.e. to reject loanwords with low support. However, the same test needs to be applied here as for the BD-based HGT algorithm, since it is unclear up to which percentage the transfer events are seen as artifacts or not. This means, different cut-off points need to be evaluated using gold-standard data in order to test the effect of support values as filters.

5.1.3 Minimal Lateral Networks

The minimal lateral networks (MLN), originally introduced by Dagan and Martin (2007) to detect gene transfers, were adapted into CHL by Nelson-Sathi et al. (2011) for the task of automated borrowing detection. The algorithm was applied to several datasets, including Indo-European (Nelson-Sathi et al., 2011; List et al., 2014a), Chinese dialects (List et al., 2014b; List, 2015), and Austronesian (Jäger and List, 2018).[12] In these studies, the MLNs are used to model language contact and to estimate the frequency and distributions of the languages under consideration.
The MLN algorithm uses the technique of gain-loss mapping to presence-absence patterns to detect lateral components (List et al., 2014a). The presence-absence patterns are obtained from cognate sets, i.e. lexical items in the basic vocabulary lists are clustered into cognate classes to compile the binary presence-absence pattern needed for the MLN analysis. The algorithm needs a reference tree, i.e. language tree, as an underlying structure to reflect the evolution and classifications of the languages. For all concepts in the dataset, specific gain-loss scenarios are inferred according to the structure of the language tree. Each scenario provides a possible explanation of the evolution of a given character along the language tree as a process of gain-loss events. One gain-loss scenario should contain only one gain event to confirm the assumption that the given character evolves in a vertical way, i.e. inheritance only. If more than one gain event is needed to explain the evolution of the character, a lateral transfer event is inferred. The MLN algorithm selects the best gain-loss scenario out of all possible ones by using ancestral genome size distribution, a key argument in the phylogenetic implementation of the algorithm. Genomes are physical entities whose inferred ancestral size can be brought into agreement with the distribution contemporary genome sizes using statistical methods (List et al., 2014a). However, the linguistic data consists of lexical items from basic meaning lists. List et al. (2014a, p. 147) restate the criterion in such a way that "those scenarios in which the number of words used to express specific meanings does not differ much between ancestral and contemporary languages." The adapted

[12]The MLN algorithm is implemented in the software package *lingpy* (List et al., 2017).

algorithm for computing MLNs for linguistic data uses weighted parsimony for the task of gain-loss mapping. The resulting network consists of the underlying language tree as phylogeny, and lateral transfer is indicated by reticulate edges. MLNs are mainly used to model language contact. The reticulate edges can, therefore, be weighted according to the weights of the gain-loss scenario to reflect the frequency of occurrence of the inferred link. The frequency of occurrence indicates the intensity of the contact between two languages.[13]

Up to now, MLNs are used to model language contact. However, the method can also give insights into the underlying lexical flow causing the contact scenario. In order to identify loanwords, the algorithm needs to be applied to single concepts instead to the whole database. For each concept in the dataset, the inferred cognate sets are used to construct binary presence-absence patterns, out of which gain-loss scenarios are computed and an MLN is reconstructed. In other words, for each concept in the database, a single MLN is computed in order to reflect the lateral transfers within a single concept, i.e. loanwords can be detected. Since reticulate edges are inferred between two languages, the algorithm can give insights into the RL and the SL of the transfer. However, the direction of transfer cannot be determined, i.e. it is not clear which language is the RL and which the SL. The reticulate edges can be seen as contact events where, theoretically, both directions of transfer are possible. This might not be a problem for language contact, however; in case of loanword detection, a clear identification of the RL and therefore of the loaned word remains, i.e. a workaround solution is presented to evaluate the algorithm statistically against the gold standard. Nevertheless, MLN is the best explored network method in linguistics and serves as state-of-the-art for modeling contact scenarios. It is worthwhile to test the MLN approach in terms of loanword detection to determine the performance of the algorithm on concept data.

5.1.4 Phylogenetic Lexical Flow Inference

The linguistic method of phylogenetic lexical flow inference (PLFI), developed by Dellert (2019a), uses information-theoretic causal inference to model language contact between languages. The toolchain contains the clustering of lexical items into cognate sets from basic vocabulary lists in the first step. In the next step, the genetic relatedness of languages and language contacts are modeled to explain how languages have influenced each other during their evolution.

The PLFI algorithm uses the overlap of cognate sets to reconstruct ancestral states

[13] See List et al. (2014a) for a graphical illustration of an example of the gain-loss mapping and an MLN network for linguistic data.

in a given language tree, and determines the lexical transfer between languages. The cognate classes are obtained from basic vocabulary lists using an internally implemented approach for cognate clustering. The *Information-Weighted Sequence Alignment* (ISWA), a newly established phonetic form alignment method, is a variant of sequence alignment to cluster cognates. The main advantage of this method is the consideration of varying information density of lemmas to align only the stems of a word. The idea is illustrated on an example given by Dellert (2019a): the alignment of the English word *freeze* and its German equivalent *gefrieren* shows that the additional material in the German word affects the alignment. In order to improve the alignment, only the stems of the words are aligned, i.e. the German word *gefrieren* is stripped into *-frier-* for the alignment with *freeze*. In addition, the word length is normalized to correct for possible effects of phoneme inventory size. The ISWA algorithm achieves a better approximation of the alignment and its corresponding alignment score, resulting in a better estimation of cognacy. According to Dellert (2019a), the method refines the established methods for automated cognate clustering and shows its strength in cross-family cognate detection.[14]

The overlap of cognate sets is the basis for the information measure for sets of languages. This, in turn, results in a measure of conditional mutual information which determines a notion of lexical transfer. The measure of conditional mutual information between languages "quantifies how much of the lexical overlap between two groups of languages can be accounted for by transmission through a third set of languages" (Dellert, 2019a, p. 4). The lexical material of the transfer between two languages is described by a path connecting the languages, which explains the overlap of cognate sets. Standard causal inference algorithms are then being applied to the resulting mutual information to determine the links that are minimally necessary to explain the pattern of lexical overlaps.

The standard causal inference algorithm (PC algorithm) can only be applied if ancestral languages are explicitly modeled. This requires a reconstruction of the cognate classes at the internal nodes of the language tree which represent the proto-languages. In the PLFI algorithm, the proto-languages are modeled using ancestral state reconstruction methods from bioinformatics. The model therefore requires a language tree as guide tree, where the proto-languages are reconstructed as sources of overlaps represented by the presence or absence of each cognate class at each node. Since causal inference is designed to extract information about the directionality of the lexical transfer between two languages, the PLFI algorithm adds directed reticulate edges to the language tree to represent the lexical transfer. The resulting network displays the minimal number of lateral connections determined

[14]The ISWA method is embedded in the PLFI toolchain and not separately available.

by the PLFI algorithm.[15]
The PLFI algorithm determines the RL, the SL, and the direction of the transfer. The information of the lexical items present in each lexical flow can be extracted from the algorithm to obtain the information of the loaned words between languages.[16]

5.2 Fundamental Requirements for Loanword Detection

All algorithms introduced for the task of automatic loanword detection are based on different pre-processing steps needed for the computation. In order to maintain a reasonable evaluation, the requirements for all algorithms should be equal to the extent possible.
The first requirement is an underlying language tree, which is needed by all four approaches to detect conflicts which are due to transfers. The two reconciliation algorithms require a concept tree and the corresponding tree replicates for the bootstrap analysis. The analyses of the MLN and PLFI algorithms use cognate classes in order to identify the classes whose evolution does not match the given language tree. All trees (i.e. the language tree, the concept trees, and the tree replicates) need to be rooted to reflect the evolutionary relationship of the languages in a hierarchical tree diagram.

5.2.1 Language Tree

The language tree represents the genetic classification of languages according to the comparative method. The tree can either be obtained from databases like Glottolog (Hammarström et al., 2018) using expert classifications, or from automatic reconstruction methods. From a linguistic point of view, the expert tree from Glottolog would be the tree of choice. The language classifications of the Glottolog tree are expected to be accurate, which can lead to better results in terms of loanword detection. The expert tree from Glottolog is multifurcating and reflects the assumption that a language can have more than two descendants. The tree reflects the clustering of the languages according to the theory obtained from classical historical linguistics. On the other hand, several linguistic studies in CHL (see e.g. Steiner et al. (2011), Jäger (2013a), and Jäger (2013b)) showed that an automatically reconstructed

[15] See Dellert (2019a) for a detailed explanation of the underlying methodology and the algorithm itself. In addition, graphical illustrations of PLFI networks are provided.
[16] Up to now, the algorithm is not available online. The results were kindly provided by Johannes Dellert, who also extracted the information of the loanwords causing the lexical flow between the languages.

bifurcating tree is a suitable alternative to an expert tree. It is therefore justifiable to choose an automatically reconstructed tree for analysis.

For the task of loanword detection, the choice of the language tree depends on the applicability of the algorithms. Three algorithms (the TC, MLN, and PLFI) can handle both binary and multifurcating trees, whereas the BD-based HGT algorithm can only process binary trees. In their study, Boc et al. (2010b) tested the BD-based algorithm on simulated nonbinary trees. However, this function is not accessible in the publicly available versions of the algorithm. In order to achieve a reliable comparison between all four algorithms, an automatically reconstructed binary language tree is preferred. Additionally, an automatically reconstructed tree reflects the language relationships according to the basic vocabulary list used for the other analyses. On the basis of the applicability of the binary tree and the valid evidence that an automatically reconstructed tree is a suitable alternative, an automatically reconstructed binary tree is the tree of choice.

In linguistics, Bayesian methods are the most popular methods to reconstruct language trees and to model language classifications (Atkinson and Gray, 2006; Greenhill et al., 2010; Gray and Atkinson, 2003). Among others, Boc et al. (2010a) used the inferred tree from Gray and Atkinson (2003) as an underlying language tree for their analysis. The framework for Bayesian inference is introduced in chapter 3.1.2, and the detailed settings for MrBayes are described in chapter 4.2. The language tree is reconstructed using the 200 most stable concepts from NELex for all languages in the dataset (Dellert and Buch, 2018).[17] The character matrix is created using the automatically inferred cognate classes from the OnlinePMI program by Rama et al. (2017). MrBayes reconstructs a tree sample according to the posterior distribution next to the consensus tree summarizing the tree sample. Since the consensus tree can result in a multifurcating tree, the maximum clade credibility tree from the tree sample selected. The MCC tree is the best representative and an actual, fully resolved binary tree from the tree sample (Heled and Bouckaert, 2013). To test whether the MCC tree is a good representative of the language classifications, it can be compared to the Glottolog tree using the GQD, introduced in chapter 4.1.4. The small GQD of $d = 0.0307$ indicates their agreement.[18] This underlines the assumption that an automatically reconstructed binary tree is a suitable alternative to an expert tree. The tree obtained from MrBayes serves as input to all loanword detection algorithms requiring a language tree.

[17] Due to the long running time of MrBayes, the number of concepts is reduced.
[18] The automatically reconstructed language tree is displayed in appendix A.2.

5.2.2 Concept Trees and Tree Replicates

The BD-based HGT and the TC algorithms require concept trees and their replicates for the reconciliation process. The establishment of the concept trees using different reconstruction methods is introduced in chapter 4. The analyses are carried out using the trees from the optimal methods determined by the MCC evaluation.
The distance-based concept trees are computed by the PMI-based NW algorithm and reconstructed using FastME. The tree replicates derived by the noisy bootstrap method are reconstructed using the same requirements. All trees reconstructed by the FastME algorithm are unrooted binary-branching trees.
Due to the different frameworks to reconstruct character-based concept trees, the analyses result in two optimal methods, one for the maximum likelihood framework, and one for Bayesian inference. The trees computed by IQTree are based on character matrices generated by using the combined approach of PMI-based NW and n-grams. IQTree results in a binary maximum likelihood tree and a sample containing binary tree replicates obtained from the implemented bootstrap analysis. The concept trees for Bayesian inference are computed using the character matrices generated from the PMI-based NW approach. MrBayes results in a consensus tree summarizing the posterior distribution of trees from the analysis. Since the consensus tree is multifurcating, the concept tree and tree replicates are obtained from the binary trees of the posterior distribution. The MCC tree is selected as concept tree to express the best representative of the whole tree sample.
The three methods result in tree samples of different sizes. For a fair comparison of the reliability of the transfer events, the tree samples should all be of the same size. The noisy bootstrap method creates a tree sample of size $k = 1,000$, which is advisable to account for the small variance of the matrices. IQTree uses the standard bootstrap method to generate the tree sample. The traditional data sample technique includes the replication of the complete initial analysis, which is computationally intensive and results in a long running time. Since the size of the tree sample can be set manually, a size of $k = 100$ is chosen to result in a reasonable running time. MrBayes determines the size of the posterior distribution in accordance with the input data, the evolutionary model, and a presumed prior distribution of trees. The tree sample can therefore be of varying size. To ensure for the same size of the tree samples, all tree samples are equated to the size of $k = 100$. From the noisy bootstrap replicates, 100 trees are chosen randomly to create the sample of size $k = 100$. For the posterior distribution of MrBayes, the trees are chosen after rejecting the first 25% of the trees in the sample. The first 25% of the trees are used for the so called *burn-in* period required by the MCMC approach. During this period, the process settles into its equilibrium distribution (Felsenstein, 2004). After this

period, MrBayes records a tree every S steps, where $S = 500$. From the remaining sample, 100 trees are chosen randomly to create a tree sample of size $k = 100$.

5.2.3 Tree Rooting

In order to achieve a hierarchical tree diagram to represent the evolutionary relationship between languages, all trees are rooted according to the same technique. Most of the standard reconstruction algorithms infer unrooted trees with branch lengths, so do the methods introduced in this thesis. Tree rooting needs to be performed in a separate post-processing step. The most popular methods for tree rooting are *outgroup rooting* and *midpoint rooting*. In a recent study, Tria et al. (2017) introduced the *minimal ancestor deviation* (MAD) rooting. Jäger (2019) compares the three automatic rooting methods using phylogenetic trees inferred from linguistic data. In his study, Jäger (2019) reconstructs language trees within the maximum likelihood framework using the ASJP database. The resulting trees are rooted using the three automatic methods, and compared to the corresponding Glottolog tree via GQD.

Outgroup Rooting The simplest way of rooting an unrooted tree is either by declaring one of its nodes to be the root, or by inserting a new root to one of the edges. A popular way in practice is to determine the root by choosing an *outgroup*. The outgroup is defined as taxa which are closely related to the ingroup, but lie outside of it and are placed on the edge, leading to this outgroup (Huson et al., 2010). To root a tree using an outgroup, prior knowledge of the language grouping is needed to classify the phylogenetic relations between the outgroup and the ingroup. Figure 5.6 illustrates the Germanic languages extracted from the language tree in figure 3.2. In this language sample, English would be defined as outgroup, since it is closely related to the main group, but lies clearly outside of it. Therefore, the root is placed on the edge between English and the remaining Germanic languages. However, considering the Germanic and Romance languages as in figure 5.7, the specification of an outgroup is unclear. In order to root this tree on an outgroup, other languages can be added to serve as the outgroup for the rooting. The insertion of a new language group to root the tree depends on the data sample. The outgroup needs to fulfill certain restrictions, like being closely related to the ingroup while still lying outside of the sample. The specification of an appropriate outgroup can therefore be problematic for some language samples. In addition, the more languages the tree contains, the more difficult it is to determine an outgroup for tree rooting. In the study of Jäger (2019), the language tree rooted by an outgroup shows a poorer agreement with the expert tree compared to the other two methods.

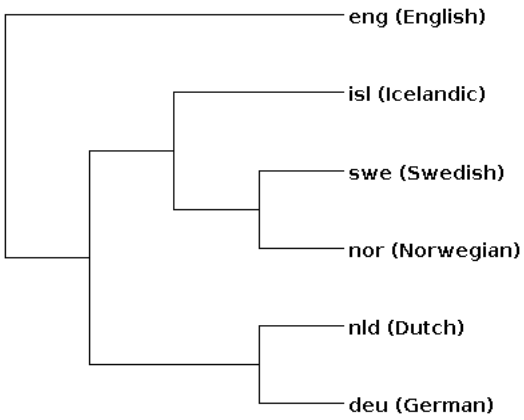


Fig. 5.6.: The language tree of the Germanic languages, based on an excerpt of the NELex database.

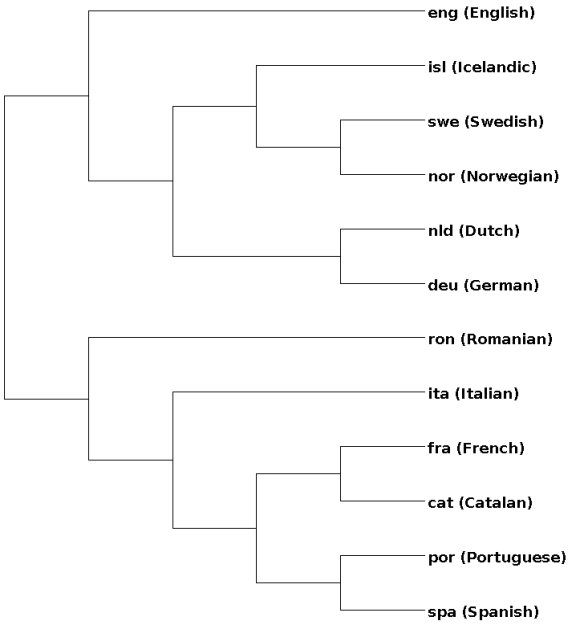


Fig. 5.7.: The language tree representing the language classification of the selection of Germanic and Romance languages.

Midpoint Rooting The midpoint rooting technique, introduced by Farris (1972), can be used to assign a root in the case where no suitable outgroup can be determined. The method locates the midpoint of the longest path between all languages in the phylogeny, i.e. it will join the most dissimilar languages to root the tree (Vandamme, 2009). The path length between two languages is the sum of the lengths of the intervening branches. The longest path is selected to root the tree. Midpoint rooting assumes that all branches have roughly similar evolutionary rates. It is advisable to

be cautious to use midpoint rooting as the method of choice, since this assumption does not hold for many linguistic datasets. In the comparison of the rooting methods, midpoint rooting outperforms the outgroup rooting. However, the comparison to the expert classification shows a small improvement of the MAD rooting method (Jäger, 2019).

MAD Rooting The minimal ancestor deviation method, introduced by Tria et al. (2017), operates on binary unrooted trees with branch lengths. No outgroup or other prior knowledge of the data under consideration is needed. In contrast to the midpoint rooting, MAD uses the idea of similar evolutionary rates instead of assuming them to be equal. Additionally, the root position is estimated by considering all branches as possible roots (Tria et al., 2017). Using this assumption, the induced ancestral–descendant relationships of all nodes in the tree can be derived and evaluated to determine the best root position. The best candidate for the root position is the branch that minimizes the relative deviations among all mean relative deviations calculated for all possible root positions (Tria et al., 2017).[19] Compared to the other rooting algorithms on language data, the MAD method performs best and is substantially better than midpoint rooting (Jäger, 2019).

The correct rooting of the trees is essential for the BD-based HGT algorithm to determine transfer events via tree reconciliation. A misplaced root can lead to false positive and false negative HGTs, since the whole reconciliation process will be affected (Boc et al., 2010b). The specification of the root position by the MAD algorithm determines the ancestral relationship of all nodes in the tree. The certainty whether the rooting algorithm inserts the root at the correct position can only be estimated with difficulty. It is unclear to which extent the resulting rooting depends on the data and the phylogenetic inference methods. Nevertheless, rooted trees indicate the evolutionary relationship between the languages under consideration and are required by the BD-based HGT algorithm. Therefore, the language tree, the concept tree, and all tree replicates are rooted using the MAD algorithm by Tria et al. (2017).

[19] See Tria et al. (2017) for a detailed mathematical description of the algorithm and experiments on biological datasets.

5.2.4 Cognate Classes

The MLN and PLFI algorithms require cognate classes to identify transfer events in the underlying language tree structure. Since NELex does not contain expert cognate judgments, the cognate classes need to be inferred automatically. The MLN algorithm requires the lexical items with their corresponding cognate classes as input. Consistent with the other analyses, the automatically inferred cognate classes from the OnlinePMI program (Rama et al., 2017) are used for the inference of the gain-loss scenarios. The toolchain of the PLFI algorithm includes the ISWA algorithm for the clustering of cognate classes used for the inference of lexical transfers. Hence, the internal cognate classes determined by the PLFI algorithm need to be used for the detection of transfer events.

5.3 Loanword Detection Methods: an Overview

In this chapter, four different algorithms are introduced for the task of automated loanword detection. The following overview should ensure for a better transparency of all the methods which can be applied to determine loanwords. The methods are summarized in table 5.1.

Algorithm	Input	Output
BD-based HGT	CT_{dow} $CT_{ML+ngramsNW}$ $CT_{Bayesian+NW}$	Loanword information from 1,016 concepts: RL, SL, transfer direction, and reliability values.
Tree comparison	CT_{dow} $CT_{ML+ngramsNW}$ $CT_{Bayesian+NW}$	Loanword information from 1,016 concepts: RL, SL, transfer direction, and reliability values.
MLN	Cognate Classes OnlinePMI	Loanword information from 1,016 concepts: 2 languages involved in the transfer.
PLFI	Cogante Classes ISWA	Loanword information from 1,016 concepts: RL, SL, and direction of transfer.

Tab. 5.1.: Overview of the methods for automatic loanword detection.

BD-based HGT and TC Algorithm The BD-based HGT algorithm requires a language tree, a concept tree, and concept tree replicates. The language tree is obtained from MrBayes, as explained in section 5.2.1. The concept trees and their corresponding replicates are the results of the MCC evaluation to determine the optimal method, introduced in chapter 4. Since there is one optimal distance-based method and two optimal character-based methods, this results in three analyses using the BD-based HGT algorithm and three analyses using the tree comparison algorithm to detect

horizontal transfer. The results contain the RL, the SL, the direction of transfer, and reliability scores for the transfer events.

MLN The MLN algorithm requires the language tree computed by MrBayes as guide tree (see section 5.2.1). The cognate classes which are used to model the gain-loss scenarios along the tree are the ones inferred by the OnlinePMI algorithm (Rama et al., 2017). For each concept, the algorithm is applied to capture single transfer events due to lexical borrowing. The result is a transfer path between two languages where the RL and SL cannot clearly be specified, i.e. no transfer direction is inferred. Additionally, no reliability values for transfer events can be established by the algorithm.

PLFI The PLFI uses the MrBayes language tree (see section 5.2.1) as guide tree to model ancestral language reconstruction and determine the direction of transfer from the overlapping cognate sets. The cognate classes are inferred internally using the ISWA algorithm, which was established along with the specific task of lexical flow inference. The information of the transfer events for lexical items can be extracted from the analysis. The results contain the RL, the SL, and the direction of transfer. Reliability values for the transfer events cannot be obtained from the analysis.

6 Evaluation and Discussion of Loanword Detection

The task of loanword detection can be carried out by four algorithms: the BD-based HGT algorithm, the TC algorithm, MLN, and PLFI. The four algorithms differ in both the underlying methodology and the results. List (2019) divides different algorithms for borrowing detection into groups. The MLN approach belongs to the group of *phylogeny-based approaches to borrowing detection*. The fundamental idea is the evolution of cognate traits along a given language tree to detect which traits are in accordance with the phylogeny, and which are causing conflicts. The ones causing conflicts in their evolution along the tree are assumed to be borrowed. The PLFI can be added to this class, since it uses cognate classes and ancestral state reconstruction to identify the overlapping cognate sets causing lexical flow in a given language tree. The BD-based HGT and the TC algorithm belong to the class of *sequence-based approaches to borrowing detection*. The basis of these algorithms are word similarities, which are used to detect borrowings. The reconstructed concept trees reflect pairwise word similarities and their corresponding clustering. The tree reconciliation algorithms estimate the discordance between the language and an individual concept tree to detect transfer events, i.e. loanwords. Since the algorithms use different methodologies to determine loanwords, a direct comparison between the algorithms needs to be treated with caution.

The results of the four different algorithms for the task of loanword detection are evaluated against a gold standard, i.e. expert judgments of loanwords. In order to evaluate the different results of the algorithms, the gold standard should contain not only the information whether a word is loaned or not, but also additional information like their donor language and the direction of the transfer. This information can only be provided by two databases, which are combined for this evaluation task.

First, the gold standard is explained in more detail to get insights into its uniqueness compared to the other available gold standards. Second, a statistical evaluation is carried out for all algorithms using a precision–recall computation. In an additional evaluation, the usage of reliability values for transfer events to reject arbitrary events is discussed for the BD-based HGT and the tree comparison algorithm. Last, the transfer direction inferred by the BD-based HGT and the PLFI algorithms are qualitatively evaluated and analyzed.

6.1 Gold Standard

The gold standard is a combined dataset consisting of NELex (Dellert, Jäger, et al., 2017) and WOLD (Haspelmath and Tadmor, 2009b). WOLD cannot be used as the underlying database, since its languages' diversity is obstructive for the concept tree reconstruction and loanword detection algorithms. The information on loanwords, however, can be used as part of the gold standard. All loanwords where the meaning, the languages, and the donor languages are included in the NELex database are added to the evaluation set. In total, 443 words are extracted from WOLD to be included in the gold standard. The NELex database is under constant development, and within the ongoing EtInEn project, expert cognate and loanword judgments are collected (Dellert, 2019b). A status of either *loaned* or *inherited* is assigned to each lexical item in NELex. Depending on the status, the ancestral or donor language is annotated, as well as the source word or the reconstructed word form, if present in the source text. For loanwords, this implicitly includes information on the borrowing process and its reconstructed history, if available. The preliminary version, kindly provided by Johannes Dellert, contains a total of 9,500 words, among which 1,200 are loanwords and 8,300 inherited words.

The preliminary gold standard provides a good distribution over the language families in NELex. For the Uralic language family, the works of Sammallahti (1988) and Janhunen et al. (1981) provide information on various subfamilies, like e.g. Finnic, Hungarian, and Permian. Etymological information for the Saami language family is extracted from Lehtiranta (2001) for 1,960 words, for the Samoyedic language family from Janhunen (1977) for 649 words, for Finnic words from Itkonen and Kulonen (1992), and for Hungarian words from Zaicz (2006). Word annotations for the Turkic language family are extracted from Sevortyan (1974) and Dybo (2013). The Mongolic words are annotated using the information present in Sanžeev et al. (2015). The Indo-European language family is represented in the gold standard by the subfamilies of Albanian (Boretzky, 1975; Orel, 1998), Balto-Slavic (Derksen, 2008; Derksen, 2014), Germanic (Kroonen, 2013), and Italic (De Vaan, 2008; Meyer-Lübke, 2009). For the Italic language family, the highest amount of words (in total 1,737) are annotated, which is of great interest in terms of borrowing and loanword detection, next to the 1,276 annotated words for the Germanic language family. The research on language contact containing Romance and Germanic languages is well studied and therefore highly suitable for the evaluation of the loanword detection algorithms in terms of donor languages and direction of transfer. In addition, the 475 annotated Albanian words contain a high proportion of loanwords, which makes them interesting to study.

The gold standard obtained from this detailed annotation contains the status of the words, i.e. loaned or inherited, plus additional information on the ancestral or borrowing history. In comparison, IELex and ASJP only provide binary presence-absence information on loanwords. In a binary annotation scheme, the loanwords can clearly be identified, whereas the absence state indicates either inherited, unknown, or missing annotation. In addition, both databases do not provide information on the corresponding donor language of the loanwords, i.e. neither the donor languages nor the direction of transfer determined by the algorithms can be evaluated. The identification of donor languages and the determination of the transfer direction, however, are the main advantages of the tree reconciliation algorithms and PLFI. This can only be obtained by the detailed and unique annotation of the gold standard developed for NELex and WOLD. It is therefore the only gold standard which can be used to evaluate all results of the automatic loanword detection algorithms, i.e. loanwords, donor languages, and direction of transfer.

6.2 Statistical Evaluation

The automatically detected loanwords by the algorithms are evaluated against the gold standard using a precision–recall analysis. In this evaluation procedure, precision is the proportion of annotated loanwords in the gold standard among the detected loanwords by the algorithms, while recall is the proportion of the total number of loanwords that were actually detected by the methods. A high precision score indicates that every loanword identified by the analysis was relevant; however, the score cannot show whether all relevant loanwords are detected. A high recall score demonstrates that all relevant loanwords were found by the algorithm, but it cannot display the amount of erroneously detected loanwords. In most cases, the two measurements are combined into a single one: the F-measure. The F-score is the weighted harmonic mean of precision and recall, which measures the accuracy of the algorithms in terms of loanword detection.

For the precision–recall evaluation, the results obtained by the algorithms are combined with the gold standard. Since the gold standard is only available for an excerpt of NELex, the algorithms' results are adapted respectively. One advantage of the gold standard is the clear distinction between inherited words and unknown or not annotated words. Precision and recall are therefore computed using a binary representation of the data, where 1 means loaned and 0 inherited. This is the basis for a clear evaluation in order to obtain a realistic estimation of the accuracy of the algorithms.

It might be the case that the BD-based HGT, the MLN, and the PLFI algorithm determine an inner node, i.e. representatives of ancestral languages, as recipient language. Due to the underlying methodologies of the algorithms, it can be assumed that the loaned word is passed on to the descendants present at the leaves. An established ancestral language is represented as set of the corresponding leaves. In the evaluation task, it is assumed that all languages in the sets are RLs. Linguistically, there is a wide range of different scenarios how a word can evolve if it was loaned by an ancestral language. On the one hand, this is highly dependent on the status of the loan, the integration, and the adaption of the word embedding at the time of inheritance. This indicates that a detailed clarification is only possible with expert knowledge. On the other hand, the classification and reconstruction of ancestral languages are based on expert judgments; so is the identification of loanwords. There are still uncertainties present which cannot be taken into account in a statistical analysis.
Since the algorithms are divided into two groups depending on their underlying methodology, the algorithms of each group are evaluated first before an overall comparison is discussed.

6.2.1 Evaluation of the Phylogeny-Based Approaches

The MLN and PLFI algorithm belong to the group of phylogeny-based algorithms. Both methods use a guide tree to detect cognate classes which are causing conflicts in the given phylogeny and are therefore assumed to be due to borrowing. From the PLFI, the loanwords and their corresponding RL can be easily extracted, since the algorithm infers directed reticulate edges between the RL and SL. In the MLN approach, the RL and SL can be detected but not distinguished, since no directionality is inferred, i.e. both directions of transfer are possible. A workaround solution is needed in order to evaluate the results of the MLNs against the gold standard.

Workaround Solution for MLN Evaluation For each detected transfer event between a language pair, the recipient language used for the precision–recall computation is chosen randomly. This sampling procedure is repeated $k = 100$ times, resulting in a set of size $k = 100$ of all precision–recall scores, out of which the mean is computed.[1]

The results of the statistical evaluation of the MLN and PLFI method are displayed in table 6.1. By using the mean values as representatives, an approximation of

[1] Different sampling sizes are tested and evaluated. The results can be found in appendix A.3.

the performance of the MLN algorithm in terms of automated loanword detection is achieved. A direct comparison of the two algorithms needs to be treated with caution.

Algorithm	Precision	Recall	F-score
PLFI	**0.179**	0.141	0.158
MLN	0.145	**0.567**	**0.231**

Tab. 6.1.: Precision–recall values for the phylogeny-based approaches to loanword detection.

The F-score gives insights into the accuracy of the algorithms, indicating that both algorithms perform poorly when it comes to loanword detection. The MLN approach is most successful, achieving the highest recall at a precision of 14%. In terms of accuracy, the MLN approach results in a considerably higher F-score compared to the PLFI approach, which is due to the high recall. The high recall retrieved by the MLN algorithm indicates the amount of annotated loanwords, while not measuring the erroneous ones that are also contained in the results. The low precision scores in both methods indicate the quality of the found loanwords in terms of the gold standard, i.e. both methods are not able to detect a high amount of true loanwords. Of course, there is an inverse relationship between precision and recall, i.e. one cannot be increased without decreasing the other. In the evaluation of the MLN, the high recall comes with the cost of the low precision value, i.e. the algorithm detects 50% of the relevant links, only 14% of the detected loanwords are correct. Overall, the MLN approach achieves better results in terms of loanword detection than the PLFI algorithm.
From a linguistic point of view, this result was not expected. The PLFI approach was developed within a linguistic framework using linguistically motivated parameters, like estimated cognate classes and reconstructed proto-languages for the detection of lexical flow. Hence, it was assumed that this additional information leads to a better estimation of lexical transfer, and therefore also to a more accurate detection of loanwords. However, the difference of 8% in the F-score shows that the opposite is the case.

6.2.2 Evaluation of the Sequence-Based Approaches

The BD-based HGT and the TC algorithm belong to the group of sequence-based algorithms detecting loanwords on the basis of word similarities. The word pair similarities within one meaning are represented in the concept trees. Both algorithms use a language tree for tree reconciliation with concept trees to detect transfer

events. The results obtained by the algorithms are the RL, the SL, and the direction of transfer, i.e. the RL and loanwords can clearly be identified by the algorithms. Since the concept trees are obtained by three different methods, i.e. one distance-based and two character-based methods, precision–recall is computed for all possible tree reconciliations to achieve an overall evaluation. The results of the statistical evaluation against the gold standard are given in table 6.2.

Algorithm	Concept trees	Precision	Recall	F-score
TC	$CT_{Bayesian+NW}$	0.153	0.430	0.226
TC	$CT_{ML+ngramsNW}$	0.147	0.395	0.214
TC	CT_{dow}	**0.165**	0.345	0.223
HGT	$CT_{Bayesian+NW}$	0.137	0.824	0.2349
HGT	$CT_{ML+ngramsNW}$	0.137	0.920	0.2391
HGT	CT_{dow}	0.137	**0.924**	**0.2398**

Tab. 6.2.: Precision–recall values for the sequence-based approaches to loanword detection.

The algorithms' accuracy, measured by the F-score, lies around 22% for all analyses. Since a high F-score implies a high accuracy, the results indicate an overall worse performance of the algorithms for the task of loanword detection. The BD-based HGT algorithm achieves the highest recall at the cost of a low precision and resulting in the highest F-score of all methods. The different methods for the reconstruction of the concept trees have no impact on the tree reconciliation and therefore also on the detection of the loanwords. The BD-based HGT algorithm shows an over-generalization of the transfer events due to the SPR moves necessary to transform the concept tree into the language tree. Since the algorithm is based on the topology of the reconstructed trees, no further linguistic information is included in order to detect true loanwords. The discordance of the two trees causes the amount of loanwords detected by the algorithms. The tree comparison algorithm achieves a higher precision, but comes at a high cost to recall, and the F-score remains at about 22%. Similar to the BD-based HGT algorithm, the concept trees reconstructed using different approaches do not have an impact on the detection of the loanwords. Overall, the BD-based HGT algorithm using the distance-based reconstructed concept trees for tree reconciliation achieves the highest F-score in terms of loanword detection.

Even though the tree comparison algorithm was developed on linguistic data, the algorithm does not improve in terms of accuracy. This is due to the fact that the basic methodology of the algorithm uses topological tree distances to estimate the discordance between a language and a concept tree. The resulting loanwords are detected according to the trees' topologies, without using additional linguistic information. An improvement of the results according to linguistic information can

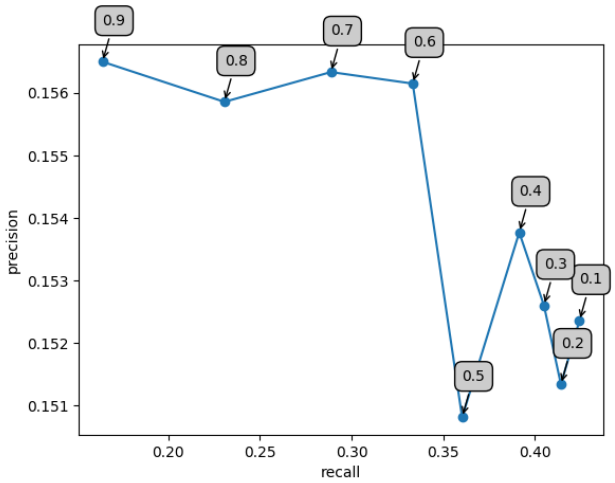
thus not be expected. The tree comparison algorithm is therefore similar to the BD-based HGT algorithm in both the underlying idea and the results for detecting loanwords.

6.3 Evaluation of Bootstrap Thresholds

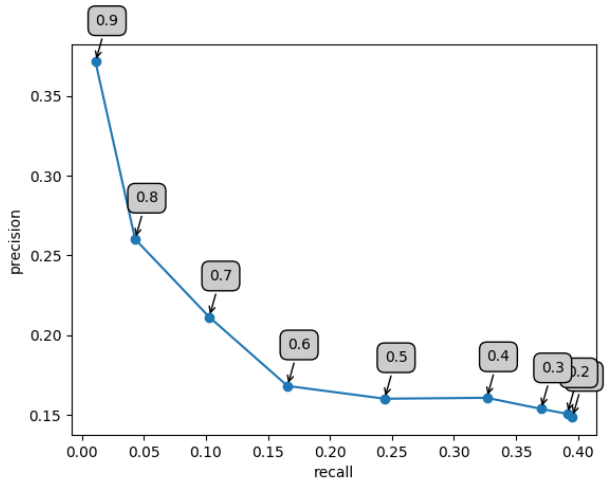
The sequence-based algorithms can verify the detected transfer events in the original scenario by using concept tree replicates to compute support values. The values indicate the frequency of occurrence of the transfer event, and can therefore be seen as a stability measurement. The BD-based HGT algorithm uses heuristic search strategies to determine the minimal number of SPR moves needed for the tree reconciliation. Therefore, the detection of exact HGT events is not possible, i.e. the algorithm can only determine potential transfer events. This can lead to the presence of arbitrary events. The tree comparison algorithm is based on the topological distance between the language and concept tree and their corresponding jackknife replicates, which is why the detection of arbitrary events cannot be excluded. This evaluation should show whether the bootstrap scores can be used as thresholds to reject arbitrary events from the results.

The bootstrap scores serve as thresholds for the statistical evaluation to exclude the events with a lower support value. For each threshold, the above explained precision–recall analysis is repeated. The results are visualized in precision–recall curves, displayed in figure 6.1 for the tree comparison algorithm, and in figure 6.2 for the BD-based HGT algorithm. The exact values including the F-scores are given in appendix A.4.

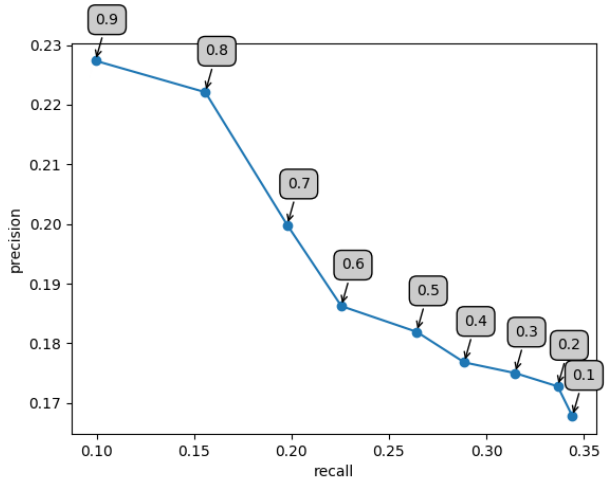
All precision–recall curves show an increasing precision at the cost of a low recall the higher the threshold gets. At a threshold of 0.9, the most accurate detected transfers are assumed, since the support values are 90% and higher. The recall increases with the lowering of the reliability values at the cost of the precision. It can therefore be assumed that the algorithms mostly detect noise. This explains their poor performance, reflected in the F-scores for the statistical evaluation in table 6.2.



(a) Precision–recall curve for the Bayesian concept trees.

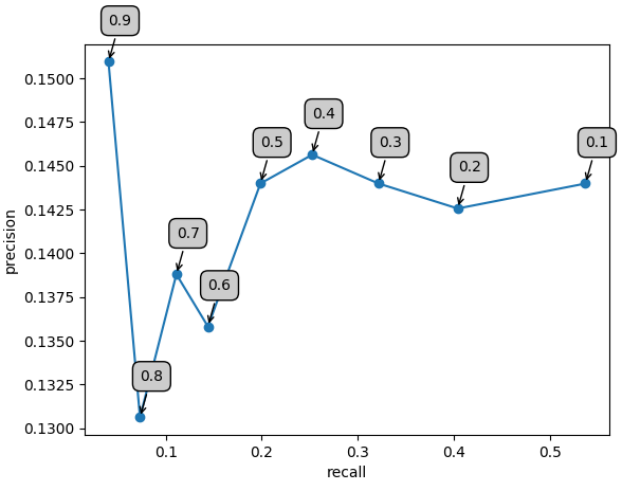


(b) Precision–recall curve for the ML concept trees.

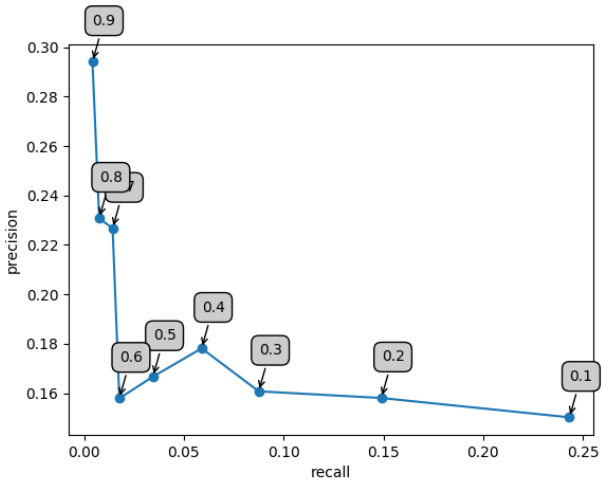


(c) Precision–recall curve for the distance-based concept trees.

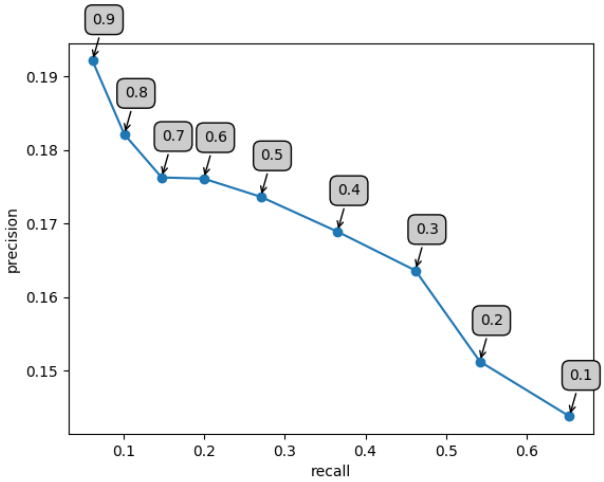
Fig. 6.1.: Precision–recall curves for the tree comparison algorithm.



(a) Precision–recall curve for the Bayesian concept trees.



(b) Precision–recall curve for the ML concept trees.



(c) Precision–recall curve for the distance-based concept trees.

Fig. 6.2.: Precision–recall curves for the BD-based HGT algorithm.

As a working sample, three words of the concept 'mountain' for three languages (English, German, and Inari Saami) are analyzed to illustrate this assumption. The English word *mountain* is clearly marked as loanword (Haspelmath and Tadmor, 2009b). The Inari Saami word *vääʹri* (Lehtiranta, 2001) and the German word *berg* (Kroonen, 2013) are inherited words from their corresponding ancestral language. In table 6.3, the detection of the three example words is visualized for both tree

Algorithm	Language	Bootstrap score	Expert judgement
HGT $CT_{Bayesian+NW}$	eng	28,7%	loanword
	smn	96%	inherited
	deu	-	inherited
HGT $CT_{ML+ngramsNW}$	eng	1%	loanword
	smn	8,9%	inherited
	deu	1%	inherited
HGT CT_{dow}	eng	12,9%	loanword
	smn	34,7%	inherited
	deu	26,7%	inherited
TC $CT_{Bayesian+NW}$	eng	-	loanword
	smn	95%	inherited
	deu	-	inherited
TC $CT_{ML+ngramsNW}$	eng	-	loanword
	smn	-	inherited
	deu	-	inherited
TC CT_{dow}	eng	-	loanword
	smn	-	inherited
	deu	36%	inherited

Tab. 6.3.: Working example: Detection of three words for the concept 'mountain' within the different versions of the algorithms.

reconciliation algorithms and all three concept tree approaches, including the reliability values for the transfers. It can be clearly seen that there is no uniform pattern between the algorithms. The BD-based algorithm detects most of the words as loanwords, which reflects the over-generalization of the algorithm, mirrored in the recall values for the complete analysis. The high recall values can be explained by the detection of a high amount of loanwords with a support value lower than 10%. Since the algorithm can only detect potential transfers with an unknown amount of arbitrary events, the underlying method using SPR for the tree reconciliation introduces a lot of noise to the analysis. On the other hand, the tree comparison algorithm never identifies the English word as loanword. It is unclear to which amount the distance measure between the language and the concept tree and their corresponding jackknife replicates introduces noise to the analysis. Both algorithms highly depend on the pre-processing steps necessary for the tree reconciliation. The

short word lengths for the computation and reconstruction of concept trees is an issue, which is reflected in this evaluation to a greater extent as in the previous evaluation.
As a result, the bootstrap values computed by both tree reconciliation algorithms cannot be used to filter out arbitrary events from the results. It is neither possible to determine a suitable threshold, nor to boost the analysis in terms of a more accurate detection rate for loanwords. However, the evaluation cannot give any insights in the source of the problem. In order to improve the results, new innovations for the subtasks and the loanword detection algorithms are needed, along with suitable datasets for training and testing. The determination of a threshold to reject arbitrary events from the results can be refined when better solutions for all involved tasks are developed.

6.4 Qualitative Evaluation of Donor Languages

The BD-based HGT algorithm and the PLFI approach are both able to determine the directionality of the transfer, i.e. identify possible donor languages. This is one of the main advantages over the other two algorithms. In order to fully understand a borrowing process, the SL and the direction of transfer are the most important factors next to the RL and the loaned word. A direct identification of a possible donor language implicitly determines the directionality of the transfer.
In the BD-based HGT algorithm, the SPR moves applied to reconcile two trees determine the direction, and therefore also the donor language. No additional linguistic information is used by the algorithm to specify the SLs. The PLFI approach identifies the donor languages according to overlapping cognate sets present at the internal nodes of the language tree. The determination of the SL, next to the directionality, RL, and loanwords, depends on the automatic cognate clustering and the ASR method. Still, the cognate information reconstructed to represent proto-languages adds linguistic information to the determination process.
Both algorithms can only identify SLs within the language sample under consideration. The analyses are based on the NELex database, i.e. all automatically detected donor languages are present in NELex. Loanwords from outside the language sample cannot be detected by the algorithms, since the determination of the directionality is limited to the underlying language tree. With this restriction, limitations for the evaluation arise. First, the loanwords borrowed from languages outside the sample need to be excluded from the evaluation. Second, not all borrowing scenarios can be identified clearly and reconstructed in classical historical linguistics, i.e. no definite

SL can be determined. These loanwords should also be excluded from the evaluation task. This results in a small amount of loanwords with an obvious borrowing scenario. The inner clades of the language tree represent the corresponding established ancestral languages. As discussed in chapter 2, the language reconstructions on an expert tree are assumed to be multifurcating, since languages can have more than two descendants. In terms of tree comparison, the automatically reconstructed language tree is equally good as their expert counterparts. However, the bifurcating tree assumes for each internal node exactly two descendants, which results in more estimated ancestral languages as assumed in multifurcating trees. However, it is still debatable which tree is the best representative for language relatedness, since the amount of proto-languages can only be estimated. For a statistical evaluation of the determined donor language, a clear mapping of the SLs provided by the gold standard to the inner clades of the language tree is essential. This cannot be guaranteed, since the donor languages in the gold standard are extracted from etymological dictionaries, and a bifurcating expert tree representing the needed proto-languages is not available.
The algorithms work on contemporaneous data, i.e. no information of the proto-languages at the inner clades is provided. The proto-languages are represented as a set of their corresponding descendants present at the leaves. There are three options of SL representations:

1. From leaf to leaf:
   fra (French) $\rightarrow$ eng (English)
2. From inner node to leaf:
   {fra, spa, cat, ita, por} (French, Spanish, Catalan, Italian, Portuguese) $\rightarrow$ eng (English)
3. From inner node to inner node:
   {fra, spa, cat, ita, por} (French, Spanish, Catalan, Italian, Portuguese) $\rightarrow$ {nld, deu} (Dutch, German)

The first example illustrates the borrowing of the French word for 'mountain' into the English language. French is detected as the donor language, and a direct link between the two languages is inferred, representing the direction. In the second example, the English word *mountain* is borrowed from an internal node representing an ancestral Romance language. The ancestral language is represented by a set of Romance languages, since the algorithms cannot define a clear proto-language. In the third example, the ancestor of Dutch and German borrowed the word for 'window' from an ancestral Romance language family represented by the set of Romance languages. In accordance with the statistical analysis, each language present in the set of recipients is identified as RL. The sets of the ancestral languages, however,

cannot be mapped clearly to the proto-languages of the gold standard. Using only the loanwords where the donor language is within NELex, each descendant from the ancestral language can be used as reference for the proto-language. This, in turn, is highly dependent on the underlying language tree and its branching. Linguistic research is concerned with the time depth of the languages and their arrangement within the language tree. There are language families, like Proto-Indo-European, with a clear, linguistically proven classification of the languages. For the Turkic language family, on the other hand, the grouping and classification of the languages is still unclear. The high similarity between the languages makes it difficult to obtain a clear clustering. Due to the different states of research, a solid statistical evaluation is impractical. It is therefore advisable to carry out a qualitative evaluation and go without a statistical precision–recall evaluation of the directionality and donor languages.

For a qualitative evaluation, three working examples are chosen out of the 1,016 concepts:

1. Concept 'mountain' with the English loanword *mountain*.
2. Concept 'egg' with the English loanword *egg*.
3. Concept 'day' with the inherited English word *day*.

For the concept 'mountain', the English loanword is a clear borrowing with a known borrowing process (Haspelmath and Tadmor, 2009b). The word was borrowed in the Middle English time period from French, replacing the inherited word *berwe/bergh* (Merriam-Webster Dictionary, 2020). There are two main reasons for choosing this example. First, the borrowed word can be identified clearly in terms of spelling and pronunciation, since it belongs to a different cognate class. Second, the grouping of the Germanic and Romance languages into two language groups is verified by experts, i.e. it is a borrowing across language borders. It can be assumed that the algorithms can detect English as RL and determine a possible donor language.
In the second example, the English word *egg* is a borrowing from Old Norse, i.e. internal borrowing within the Germanic languages. Further, it is a cognate class internal borrowing, since the replaced English word *ey* was cognate to *egg*. The PLFI algorithm is based on cognacy and overlapping cognate sets. It should not be able to detect English as RL, and can therefore not determine a possible donor language. For the BD-based HGT algorithm, it is also difficult to detected English as RL due to the high similarity of the words within the Germanic languages.
The concept 'day' is the second most stable concept in NELex (Dellert and Buch, 2018), i.e. most of the words are inherited. The English word *day* can be reconstructed to its Proto-Indo-European root and is clearly inherited (Kroonen, 2013).

Due to the high stability of the concept, the algorithms should not detect English as RL, and therefore no donor language at all.

Evaluation of 'mountain' For the concept 'mountain', the results of the determined SLs by the algorithms are displayed in table 6.4.

Algorithm	RL	SL	Reliability score
GS	eng	fra	-
HGT $CT_{Bayesian+NW}$	eng	kaz (Kazakh)	28.7%
HGT $CT_{ML+ngramsNW}$	eng	(ita, fra, cat, por, spa)	1%
HGT CT_{dow}	eng	(ita, fra, cat, por, spa)	12.9%
PLFI	eng	cym (Welsh)	-

Tab. 6.4.: The donor languages determined by the gold standard and found by the algorithms for the English loanword in the concept 'mountain'.

The gold standard specifies French as donor language, since it is a direct ancestor of Old French. The BD-based HGT algorithm determines the donor language and the directionality of the transfer in accordance to the SPR move applied to transform the concept tree into the language tree. The analysis using Bayesian concept trees for the reconciliation infers Kazakh, a language from the Turkic language family, as donor. The concept tree replicates verify this result in 28.7% of the tree reconciliations. The Kazakh word for 'mountain' is transcribed as `tau`, where the `a` and `u` should be aligned in both words. However, the remaining sounds should cause gaps in the alignment, i.e. word similarity is not the reason for the inference of the HWT. The underlying heuristic search of the BD-based HGT algorithm is responsible for this SPR move identifying the wrong donor language. For the distance-based and maximum likelihood concept trees, the algorithm determines a set of Romance languages as donor languages. The set includes other Romance languages containing related words for 'mountain' belonging to the same cognate class. Not only can the set be seen as the ancestor of French, but French can be isolated from the set, serving as correct identification of the SL. Nevertheless, the low bootstrap support values indicate the uncertainty of the algorithm, i.e. this transfer is only found by a minority of the reconciliations using tree replicates. The PLFI algorithm identifies Welsh as donor language. The transcription of the Welsh word is `m3ni8`, which should be clustered in the same cognate class with `mauntin`. Due to the automatic cognate clustering and the ASR methods used for the reconstruction of proto-languages, the detection of the correct SL was not possible. The algorithm relies on a toolchain of various tasks, which makes it difficult to identify the source of the wrong determination. Overall, two versions of the BD-based HGT algorithm correctly determine the donor language of the borrowing process. However, the

support for this transfer is so low that the detection of the borrowings could be arbitrary events due to the underlying heuristic search procedure.

Evaluation of 'egg' The results of the identified donor languages for the concept 'egg' are displayed in table 6.5.

Algorithm	RL	SL	Reliability score
GS	eng	isl	-
HGT $CT_{Bayesian+NW}$	eng	kmr (Northern Kurdish)	86%
HGT $CT_{ML+ngramsNW}$	eng	isl	1%
HGT CT_{dow}	eng	(swe, nor, dan)	95%
PLFI	-	-	-

Tab. 6.5.: The donor languages determined by the gold standard and found by the algorithms for the English loanword in the concept 'egg'.

The English word *egg* is an interborrowing from Old Norse. The gold standard specifies Icelandic as representative language of Old Norse, since Icelandic is the language closest to Old Norse. It was assumed that the detection of the English loanword *egg* would be difficult, if not impossible for the two algorithms. In the case of PLFI, the algorithm does not detect a borrowing scenario with English as RL. The algorithm clusters the English word correctly within the same cognate class containing the other Germanic languages. Since the loanword detection is based on the reconstruction and overlapping of cognate classes, this internal borrowing, where the loanword and its replaced counterpart belong to the same cognate class, cannot be found by the PLFI method. The BD-based algorithm using Bayesian concept trees identifies Northern Kurdish as SL, with a high bootstrap score indicating the frequency of occurrence within the reconciliation using tree replicates. There is a high probability that the alignment of the the Northern Kurdish word `hek` and the English word `eg` receive a high similarity score due to the alignment of `e-e` and `g-k`. This is the downside of using phonetic word alignment on short sequences. Short words can show a high agreement with each other, since the correct alignment of a small amount of sounds is sufficient to receive a high similarity score. The BD-based HGT algorithm using concept trees computed by IQTree correctly identifies Icelandic as donor language. However, the verification of the reconciliation by tree replicates results in a high uncertainty, mirrored in the low bootstrap score. The tree reconciliation with distance-based concept trees achieves the best result. An established ancestral language containing Swedish, Norwegian, and Danish could be identified as donor language, which could be interpreted as Old Norse although Icelandic is missing. The high bootstrap support indicates a clear identification of the transfer among 95% of the tree reconciliations using tree replicates. However, in

a statistical analysis, the donor language would be rejected and counted as wrong identification, since Icelandic is not contained in the set of languages. Overall, the BD-based algorithm correctly identifies the donor language in two cases, whereas the PLFI approach is not able to detect the internal borrowing.

Evaluation of 'day' Table 6.6 displays the results of the determined SLs for the concept 'day'.

Algorithm	RL	SL	Reliability score
GS	eng	-	-
HGT $CT_{Bayesian+NW}$	eng	jpn (Japanese)	11.9%
HGT $CT_{ML+ngramsNW}$	eng	che (Chechen)	12.9%
HGT CT_{dow}	-	-	-
PLFI	eng	cym (Welsh)	-

Tab. 6.6.: The findings of the algorithms for the English native word in the concept 'day'.

The English word **day** is inherited and can be traced back to a Proto-Indo-European root (Kroonen, 2013). This is correctly identified by the BD-based HGT algorithm using distance-based concept trees, where English as RL is not present in the results. The variant using Bayesian concept trees wrongly identifies English as RL and Japanese as donor language. Since English is a well-known donor language for loanwords in Japanese, the direction could be inferred wrongly. However, this assumption is not underlined by the gold standard, where the Japanese word for 'day' is not listed as loanword (Haspelmath and Tadmor, 2009b). The BD-based HGT algorithm using concept trees obtained from IQTree determines English as RL and Chechen, a language from the Nakh-Daghestanian language family, as source language. The Chechen word `de` for 'day' shows a high similarity to the English word `dei`. The transfer is therefore due to short sequence alignment. All in all, both variants show low support values for the transfer events, i.e. these could also be arbitrary events found by the algorithm. The PLFI method identifies English as RL and Welsh as SL. However, the English word `dei` and the Welsh word `di8` belong to the same cognate class, i.e. the transfer is an effect of wrong cognate clustering.

In summary, it seems that an accurate determination of donor languages along with the correct inference of directionality can only be estimated by chance. The algorithms show no clear tendency for a systematic identification of the SL in the language tree. The BD-based algorithm can only detect potential events, while arbitrary events cannot be rejected from the results and cause additional noise. The determination of the SL highly depends on the underlying language tree and

the inferred SPR moves during the reconciliation process. No further linguistic information is provided to improve the detection of RL, SL, and the inference of directionality. It is, therefore, not surprising that the correct SL can only be detected by chance. The PLFI algorithm is based on cognacy and the reconstruction of cognate sets to represent proto-languages at the inner nodes of the tree. Although the algorithm highly depends on the pre-processing steps, a better chance to correctly identify the SL due to implicit linguistic information present at the inner nodes was assumed. However, the qualitative evaluation shows the opposite. It seems the systematic of finding overlapping cognate sets to determine the SL and the direction of transfer leads also to arbitrary results, and the correct SL can only be determined by chance. Overall, both algorithms correctly identify the direction of transfer, since the RL is detected in most cases. However, the accurate determination of the SL is rather by chance than following any systematic which can be linguistically motivated. This evaluation underlines the poor performance of the algorithms in terms of loanword identification. Unfortunately, the algorithms can neither be successfully applied to the task of automated loanword detection, nor achieve a breakthrough in terms of SL identification and the inference of directionality within evolutionary networks.

6.5 Discussion of the Results

All presented evaluations reflect a rather poor performance of the algorithms for the task of automatic loanword detection. In an overall comparison, the results of the statistical evaluation are similar in terms of F-scores. For the MLN, the BD-based HGT, and the TC algorithm, the loanword detection accuracy is around 23%. The PLFI approach performs worst, with an F-score of 0.15. The BD-based HGT methods achieve the highest recall, but due to a low precision, the F-score remains at 0.23. The reliability scores computed by the sequence-based algorithms through concept tree replicates cannot be used to reject arbitrary events from the results. It was assumed that low support values are an indication for arbitrary transfer events, since the frequency of occurrence in the analyses is low. The evaluation has shown that the reliability values could neither be used to determine a threshold to sort out arbitrary events, nor to improve the accuracy of the loanword detection. The BD-based HGT algorithms and the PLFI are able to infer the directionality of the borrowing process along with a possible SL. However, both algorithms only detect the right donor language by chance. The inference of the correct directionality can merely be estimated, since this implies the correct identification of the RL. In the qualitative

evaluation using three working examples, the identification of the RL, and therefore the right direction of the transfer, was correct. However, the poor performance of the algorithms indicates that in an overall analysis the correct determination is still problematic. The rather poor performance of the algorithms can have several reasons next to the underlying methodologies used in each algorithmic framework. The main reasons are discussed in the following.

Other Evolutionary Changes A word occurring nowhere else in the dataset is indistinguishable from the effects of internal replacement, which is often due to semantic shift. For instance, the concept 'head' has two realizations in German: *Haupt* and *Kopf*. *Haupt* is the inherited word, with a figurative meaning nowadays. It is therefore not included in NELex. In earlier times, the word *Kopf* represented a vessel for drinking and was borrowed from the Latin word *cuppa* (Dudenredaktion, 2013), i.e. it is cognate to the English word *cup* (Dudenredaktion, 2013). Due to semantic shift and language evolution, the word *Kopf* has changed semantically. This type of meaning change cannot be detected by the algorithms, since *Kopf* and *cup* now represent two different concepts. The algorithms are applied concept-wise and focus on the detection of loanwords within a single concept, i.e. semantic changes during language evolution cannot be captured by the approaches. It is unclear to which extent other evolutionary processes influence the results in both the pre-processing steps needed for the analyses and the actual application of the methods. The pre-processing methods, like alignments and cognate class assignments, are applied to contemporary phonological representations and cannot take the original state of the word into account without using any expert linguistic information. If this cannot be captured and analyzed in the pre-processing steps, the information is lost and cannot be detected by the loanword detection algorithms. For example, without any additional linguistic information, the German word *Kopf* cannot be identified as loanword by the detection algorithms. Therefore, the results obtained from the loanword detection cannot give any insights whether different evolutionary processes are involved next to the process of borrowing detection. A distinction between different evolutionary processes within the computation of the algorithms is one difficulty in the procedure of automatic loanword detection.

Automated Cognate Detection For the phylogeny-based algorithms, the automated cognate judgments limit the achievable performance of the algorithms. Compared to gold standard data, algorithms for cognate clustering achieve promising results (Rama et al., 2018). However, at the description level of cognate classes, it is impossible to detect borrowings within the same cognate class. For instance, the

English word *egg* is a borrowing from Old Norse *egg*, replacing the inherited word *ey*. The English word is, therefore, annotated as loanword in the gold standard. However, this kind of internal borrowing is invisible to the phylogeny-based methods, since the loanword and the inherited word belong to the same cognate class. This demonstrates that the information loss on the description level can have a negative impact on the performance. Additionally, algorithms for cognate detection might cluster the words differently, as proposed by linguists. The varying amount of cognate classes can also have an impact on both the phylogeny-based loanword algorithms and the reconstruction of character-based concept trees. The assignment of less or even one cognate class to the words within one concept can lead to spurious concept trees, since the effect of small matrices on tree reconstructions is unclear. In contrast, a small amount of cognate classes have no effect on the analysis using phylogeny-based approaches. Linguistically, a small amount of cognate classes represent the strong relationship between the languages in the sample. If all words in a concept are cognates, the phylogeny-based approaches would not detect any loanwords in the analysis. If the amount of cognate classes is small, the algorithms would detect a smaller amount of loanwords, if any. It could be assumed that this leads to more accurate results in terms of loanword detection, since the cognate classes add linguistic information to the analysis, and the probability of detecting arbitrary loanwords should decrease. This, however, is not mirrored in the results of the evaluation presented in table 6.1. The difference in the amount of cognate classes automatically assigned to the words within one concept, compared to an expert clustering, can lead to historically incorrect reconstructions of the character evolution along the tree. This, in turn, affects the loanword identification carried out by the phylogeny-based algorithms.
Overall, the task of automated cognate detection achieves promising results which can be included in the toolchain for loanword detection. Although some effects cannot be captured by the algorithms, the linguistic information is important to distinguish inherited words from loanwords.

Short Word Length For the sequence-based algorithms, the word length can have an impact on the computation of the matrices and the reconstruction of the trees. Compared to genome sequences, the short length of the words can induce noise resulting in less reliable trees as expected from the research on gene trees. The number of sounds in a word determines the number of sequences used for the alignment analysis. Shorter sequences can lead to small distance variations between the languages within a concept. This is reflected in the small range of the distances present in the matrix. The effect of short sequences on the language clustering

obtained from the tree reconstruction methods can only be estimated. In theory, the distances in the matrix are the decision basis for the tree reconstruction algorithms to cluster the languages. This indicates that the clustering is dependent on the computed language distances. If the distances and their variations are small, it is more difficult for the algorithm to compute the clustering of the languages.
For the character-based methods, a small amount of cognate classes can lead to spurious results of the concept tree reconstruction. However, the usage of different approaches to account for word length by using subsequences does not improve the results in terms of loanword detection. It was assumed that more information of the characters leads to better character matrices, which could result in reliable concept trees. The approaches using subsequences are based on the phonological representations of the words. No further linguistic knowledge is added to the computation. Although the number of characters in the matrix increases, the underlying information does not gain on information. This, in turn, could explain the missing improvements, which were expected by the analysis. Additionally, the concept trees could only be indirectly evaluated using the tasks of loanword detection. The resampling analyses only give insights in the stability of the inner clades and the tree structure. No assessment of the tree could be made in terms of correctness. To my knowledge, no equivalent methods for gene tree evaluation are present, which is why the reliability of gene trees in bioinformatics is still questionable, and thus, also the reliability of the concept trees. The question of gene or concept tree evaluation remains, therefore, unanswered. It is purely speculative to which extent the word lengths have a negative impact on the performance of the tree reconciliation algorithms.

Tree Rooting The rooting of the trees is another blind spot in the pre-processing analyses. Both the language and the concept tree including its replicates are rooted to obtain a hierarchical clustering of the nodes. A rooted phylogenetic tree is able to represent the history of the languages and the latest common ancestor of all entities at the leaves of the tree. The rooting of the language tree is in accordance with the classical comparative method, which identifies the root along with the reconstruction of the proto-languages. In terms of concept trees, the declaration of one latest common ancestor for all words presenting the same meaning should be assessed critically. The assumption whether all words in a concept can be reconstructed to one latest common ancestor depends highly on the linguistic theory and the languages under consideration. For a language family like Germanic, this assumption can easily be verified, since linguistic research has shown that all languages are highly related to another. Additionally, they descend from another common ancestor,

namely Proto-Indo-European (PIE). PIE contains several subfamilies, where several reconstruction possibilities could occur: all words have one latest common ancestor in PIE, PIE contained more than one word for the concept, or the words emerged in the subfamilies without any representation in PIE (e.g. through borrowing, semantic change, or neologism). The uncertainty of the identification of loanwords at a specific time in the reconstruction, mentioned by Kessler (2001), also holds for the identification of inherited word forms. PIE is a reconstructed language based on the comparison of contemporaneous languages, while considering their records from earlier times if known. The reconstructed language reflects the occurrences of the words in its daughter languages, i.e. it determines both the number of words and the word form. This affects the linguistic theory about the latest common ancestor of words. However, it is assumed that all languages on a language tree descend from one latest common ancestor. Words are passed on from the mother to the daughter language, reflecting the inheritance of words forming a new language. Intuitively, the same assumption can be made for the descent of words. This is in accordance with the biological theory, where genes are passed on through sexual reproduction and therefore descend from two parents. All of these arguments indicate that the assumption of one latest common ancestor of words would be in accordance with the classical comparative method. However, an unambiguous statement whether a concept tree should be rooted by one latest common ancestor or not can therefore not be made.

The rooting affects the hierarchical representation of the tree, which can have an impact on the identification of the donor languages within a borrowing process. The BD-based HGT algorithm highly depends on the input data, i.e. the language and concept tree. The rooting of the trees has an impact on the tree reconciliation, since the transfer events correspond to the SPR moves applied to transform the concept tree into the language tree in order to identify loanwords, their SL, and the directionality of transfer. The PLFI algorithm is based on the language tree and the reconstructed cognate classes. The distribution of the languages and their ancestral states in the tree affects the ASR and the detection of overlapping cognate sets to determine loanwords, their SL, and the direction of the transfer. The rooting of the trees plays, therefore, a major role in order to achieve an accurate detection of the borrowing process for both the BD-based HGT and the PLFI algorithm.

Since the manual rooting of a linguistic tree is already a challenging task, the correctness of automatic methods is questionable. The automatic determination of the root using the MAD algorithm highly depends on the reconstructed trees. The effect of negative rooting on the loanword detection algorithms can only be estimated. Incorrect rooting has a higher impact on the tree reconciliation methods than on the phylogeny-based algorithms. A misplaced root in the language or

the concept tree can lead to the detection of false positive and false negative transfer events using the BD-based HGT algorithm (Boc et al., 2010b). In the tree comparison algorithm, the rooting can affect the language clustering, and therefore the discordance measurement using the NTD. For the phylogeny-based algorithms, a misplaced root in the language tree has an impact on the estimation of the character evolution along the phylogeny, which influences the results. However, a manual rooting of the trees would require the availability of expert trees. The automatically reconstructed language tree could be replaced by the Glottolog tree to improve the results of the loanword detection algorithms. This is, however, not yet possible for the BD-based HGT algorithm. Expert concept trees are not available, which is why the analyses need to be carried out using automatically reconstructed and rooted concept trees.

Borrowing from Outside the Sample All loanword detection algorithms can only determine loanwords within the language sample. Loanwords borrowed from languages not included in the sample cannot be identified as such. The cross-linguistic studies made in this thesis are all restricted to the available datasets and the gold standard present for the analysis. All algorithms operate on the basis of a language tree and on some form of linguistic information of the lexical items used. Linguistic information which is not included in the dataset cannot be taken into account by the algorithms, since it is not given as input. This limits the automatic detection of loanwords compared to the classical comparative method, which is based on linguistic knowledge, and additional information can easily be taken into account. This, of course, also affects the detection of donor languages in the BD-based HGT and PLFI algorithm. Both algorithms can only use the information given in the language tree to determine donor languages. Loanwords with donor languages not present in the sample, i.e. borrowings from outside the sample, cannot be captured by the algorithms. This can either lead to a wrong determination of the SL, or to undetected loanwords.

The preliminary gold standard limits the evaluation of the detected loanwords of NELex. On the one hand, this gold standard is the most precise and detailed gold standard available. A clear distinction between inherited and loaned words can be made, where unclear cases and unannotated words could be rejected from the evaluation. This leads to a better estimation of precision–recall than using the binary gold-standard data from the ASJP database. On the other hand, the annotation of such a gold standard is a time-consuming task. Linguists classify inherited and loaned words on the basis of etymological dictionaries. No attention can be paid to the limitation of the algorithms in terms of intersample borrowing detection.

Such limitations affect both the evaluation of the loanword detection and the development of suitable algorithms. It is therefore advisable to extend the development of annotated datasets in order to improve the results through achieving innovations in the development of suitable algorithms.

Interpretation of the Network The statistical evaluation is an explicit interpretation of the network, i.e. each reticulation is treated as an evolutionary event due to horizontal transfer. No further linguistic knowledge is added to the analysis. However, the interpretation of the network has a great impact on the explanations of the transfer events. Morrison (2011) argues that an implicit interpretation is crucial to distinguish the various causes of evolutionary events represented by the reticulations. In other words, without an implicit interpretation using linguistic knowledge, borrowing cannot be distinguished from other evolutionary events like semantic change; the correct history of the word and its SL cannot be determined; unclear borrowings from outside the sample cannot be identified; and arbitrary transfers cannot be identified. The graphical representation of the networks can be used as starting point for such an implicit analysis. In linguistics, the difficulty for a proficient analysis of the transfer events lies within the complexity of the borrowing process. The adaptation and integration of the words into the RL are individual, not to mention the various types of loanwords which were classified by Haugen (1950) and others. This detailed information cannot be analyzed by the loanword algorithms. Linguistic knowledge, and therefore an implicit analysis, is needed for a full reconstruction of the borrowing process. However, the identified loanwords can be evaluated using an explicit analysis, since in linguistics the two main processes, borrowing and hybridization, are clearly distinct.

The mentioned limitations to detect loanwords influence the performance of the algorithms. The question arises whether the detected loanwords are found by chance. In theory, the algorithms are suitable for the detection of borrowings, and the underlying methodologies indicate promising results. However, all algorithms highly depend on the dataset and several pre-processing steps. The pre-processing toolchain builds on established subfields of CHL, such as sequence alignment and cognate clustering. The performance of the algorithms is bound to improve further when better solutions for these subtasks are available. In addition, the limited number of suitable datasets for the different tasks of linguistic inference limits the development of applicable algorithms, since data for testing and training is missing (List, 2019). There is a crucial need of annotated datasets to test both the pre-processing and the loanword detection methods.

7 Conclusion

This thesis presents the adaptation of the tree reconciliation framework into computational historical linguistics and its application using linguistic data. The approach reconciles two linguistic trees to compute the discordance between them in order to detect horizontal transfer. The background in historical linguistics is reviewed to specify guidelines and definitions for a clear usage and implementation of the tree reconciliation approach. The most important insight is that the identification of loanwords, their adaptation process, and the detection of the SL are challenging tasks. The adaptation and integration of loanwords is language, if not process-specific. No universal assumptions and patterns can be established which could be used as additional information for the task of automatic loanword detection. There are a few linguistic constraints on borrowing, which are included in the analysis. First, the borrowing hierarchy indicates that nouns are most likely to be borrowed, which is reflected in the usage of a basic vocabulary list. Second, the classical comparative method serves as basis, since cognates and sound correspondences are criteria to distinguish inherited from loaned words. Third, the clear distinction between hybridization and borrowing leads to an unambiguous analysis of loanwords.

The underlying methodology of the algorithms requires a language and a concept tree for the process of tree reconciliation. The language tree displays the classification of the languages according to a basic vocabulary list, whereas the concept tree represents the language clusters within a single concept, i.e. the evolution of individual words. The comparison of a single concept tree to a language tree gives insights into the different evolutionary histories of the words. The mismatches between the trees are due to horizontal transfer, where the explicit identification of the loanword, the RL, the SL, and the directionality can be inferred. Language trees are a well-known model to display language classifications, and suitable methods for their automatic reconstruction are established in the field of CHL. Up to this point, less attention is paid to the reconstruction of concept trees to illustrate the evolution of individual words using the tree model. Although several studies used concept trees for the detection of horizontal transmission, the reconstruction was a secondary effect of the analyses (see Delz (2013), Delz (2014), and Willems et al. (2016)). Since the concept trees play a major role in the tree reconciliation, different approaches to compile distance and data matrices for concept tree re-

construction are introduced in chapter 4. The main challenge in linguistics, and also in bioinformatics, is to assess the quality of the concept trees. There are no expert concept trees which can be used for a tree comparison to determine the correctness of the trees. Therefore, statistical resampling techniques are used to establish the reliability of the trees. The character-based algorithms, IQ-Tree and MrBayes, already include the computation of a tree sample to determine the stability of the clades. For the distance-based method, the noisy bootstrap technique was introduced to assess the reliability of inner clades. All resampling techniques create a tree sample out of replicates, which can be used to find the optimal method for the reconstruction of concept trees. This is an important step towards an unbiased comparison of the different methods independently from other tasks like loanword detection. In summary, there are three methods for concept tree reconstruction found to be ideal in comparison to the other methods under consideration, namely CT_{dow}, $CT_{ML+ngramsNW}$, and $CT_{Bayesian+NW}$. The resulting concept trees and their replicates from these three methods are used for the loanword detection tasks.

The tree reconciliation algorithm from Boc et al. (2010b) is used to infer HWT events which result in the detection of loanwords. The algorithm is an implementation of the tree reconciliation procedure described in chapter 3.3, and can be applied to linguistic data to test whether a direct adaptation of the algorithm is fruitful for the detection of borrowings and the corresponding loanwords. The BD-based version is the one of choice used for the linguistic analyses. It provides information about single transfer events, the loaned word, the RL, the probable SL, and the direction of the transfer. These results can additionally be illustrated in an evolutionary network. The BD-based HGT algorithm is applied to each concept of NELex to extract the loanwords detected by the algorithm for the whole dataset.

The BD-based HGT algorithm is compared to three algorithms from CHL for the task of borrowing detection. The first algorithm, introduced by Delz (2014), is a tree comparison method based on the idea of tree reconciliation to detect loanwords. The algorithm was only evaluated manually, since at the time the database under consideration did not contain expert judgments for loanwords. The analysis is repeated using the data from NELex and serves as a direct comparison to the BD-based HGT algorithm. The MLN approach, originally introduced in phylogenetics by Dagan and Martin (2007), was applied to the task of borrowing detection by Nelson-Sathi et al. (2011). In contrast to tree reconciliation methods, MLNs establish the evolution of binary characters along a given language tree to detect the transfer events. In several studies, MLNs are used to model language contact (see e.g. Nelson-Sathi et al. (2011), List et al. (2014a), List et al. (2014b), List (2015), and Jäger and List (2018)). In addition, the method can be applied to concept data in order to detect the lexical items included in the contact scenario. The third linguistic

model is the PLFI model developed by Dellert (2019a). The approach makes use of information-theoretic causal inference to discover language contact and to generate an evolutionary network including reticulate edges. The detected contact situations by the PLFI method contains information on specific borrowing events and lexical items involved in the transfer which can be statistically evaluated.

The algorithms are divided into phylogeny-based (MLN and PLFI) and sequence-based methods (BD-based HGT and TC algorithm) to achieve a reliable statistical evaluation. Additionally, two further evaluations are carried out. First, the usage of reliability values of the transfer events for the sequence-based algorithms are evaluated to determine whether arbitrary transfers can be rejected from the results. Second, a qualitative evaluation on the identification of SLs is carried out for the BD-based HGT and the PLFI algorithm. All results show a rather poor performance of the four algorithms. This was expected, since the task of loanword identification is challenging and involves several intermediate steps on the way to the identification and reconstruction of a borrowing process. The detailed review in chapter 2 presents the great variety of linguistic and social constraints which led to the full integration and adaptation of the loaned words in the RL, not to mention the great challenge to reconstruct the language history, the contact scenarios, and all additional factors involved in the borrowing process. It is therefore not surprising that an automatic identification of loanwords without any additional linguistic information results in a poor performance. In addition, the identification of loanwords comes with some limitations which cannot be captured by the algorithms.

In summary, the algorithms on automatic loanword detection are still in their infancy. The BD-based HGT algorithm from phylogenetics, adapted for this task into CHL, is a step into the right direction. Using the framework of tree reconciliation, the algorithm can detect the loanword, the RL, the SL, and the directionality of the transfer. This can only be captured by the PLFI network, which, however, shows the worst performance compared to the other three algorithms. The thesis shows that a direct transfer of HGT algorithms in CHL results in an equally poor performance as the algorithms developed for the purpose of detecting horizontal transmission. Although further improvements of the methods are necessary, the detection and identification of loanwords is a first step for a deeper understanding of language contact, contact scenarios, and possible types of loanwords present in the data. The development of appropriate approaches and algorithms has been neglected in CHL for far too long. It was, therefore, worthwhile to adapt phylogenetic methods into CHL to shed light on their application, performance, and results obtained from the analyses.

7.1 Outlook

Since the task of automatic loanword detection highly depends on the pre-processing steps, there are various options to improve the performance of the tree reconciliation algorithms.

The implementation of the BD-based HGT algorithm could be extended to work with multifurcating language trees. If an expert language tree could be used for the reconciliation, more linguistic information on the languages' history is included in the computation. This could not only improve the results of the loanword detection along with the reliability of the transfer events, but also the determination of the potential donor languages.

The uncertainty of the concept trees' reliability is a far-reaching problem. No expert concept trees are available, and a reconstruction by linguists is time-consuming. The development of methods using more linguistic information on the history of words is necessary. However, this would include information on inheritance and borrowing of the words, along with an integration of possible adaptation steps of loaned words, which need to be explored in both classical and computational historical linguistics. Additionally, the inclusion of linguistic information needs to be treated carefully to avoid circularity in the process of automatic clustering and reconstruction of the languages.

The problem of short sequence alignment is crucial in terms of concept tree reconstruction, which leads to the question whether basic vocabulary lists are suitable for this task. However, the issue of data availability limits the application of other datasets. If different datasets are used for the concept tree reconstruction, it must be ensured that high-quality gold-standard data is available for the evaluation. In addition, linguistic constraints on borrowing should be considered along with the choice of the dataset using different concepts. The WOLD database includes concepts of the modern world which reflect a high amount of borrowing. However, many of these concepts are not culture-free, i.e. a high proportion of the languages have no word representing this meaning (Tadmor et al., 2010). It is obvious that the choice of the underlying dataset is limited on further restrictions which need to be considered for the task of automated loanword detection.

Each pre-processing step needed for a detailed loanword detection analysis requires more research to develop reliable methods which could be used for further processing. This includes both the development of concept tree reconstruction methods and the automatic rooting of trees. The issue of data availability for training and testing algorithms and enhancing the development of suitable methods is still present in CHL. There is a major need for high-quality databases which can serve as input for

the automatic phylogenetic methods. In addition, annotated gold-standard data is needed for a more detailed annotation of the results. The research group around NELex is working on a high-quality gold standard, which clearly had a positive impact on the evaluation of the algorithms. With an increasing gold standard, more detailed evaluations are possible to gain more profound insights into the performance of the loanword detection algorithms.

7.2 Final Remarks

The main contributions of this thesis to the field of computational historical linguistics are rather related to methodology and algorithms as to the results themselves. It has been shown that both is possible: the reconstruction of concept trees to model the evolution of words along a phylogeny, and the application of tree reconciliation to detect loanwords, the RL, the SL, and the direction of transfer. Although the methods for the task of automated loanword detection are still in their infancy, the HGT methods adapted from phylogeny open a new research field introducing new methods and challenges to CHL. It is a long way to develop a model for a complete identification and reconstruction of borrowing processes. However, tree reconciliation methods are a step into the right direction in order to model transfer events.

Bibliography

Alonso, Laura, Irene Castellon, Jordi Escribano, Xavier Messeguer, and Lluis Padro (2004). "Multiple Sequence Alignment for Characterizing the Linear Structure of Revision". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (cit. on p. 110).

Atkinson, Quentin D (2011). "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa". In: *Science* 332.6027 (cit. on p. 87).

Atkinson, Quentin D and Russell D Gray (2005). "Curious Parallels and Curious Connections – Phylogenetic Thinking in Biology and Historical Linguistics". In: *Systematic Biology* 54.4, pp. 513–526 (cit. on pp. 1, 22, 61).

– (2006). "How Old is the Indo-European Language Family? Illumination or more Moths to the Flame". In: *Phylogenetic Methods and the Prehistory of Languages* 91109 (cit. on pp. 103, 105, 138).

Baldauf, Sandra L (2003). "Phylogeny for the Faint of Heart: a Tutorial". In: *TRENDS in Genetics* 19.6, pp. 345–351.

Bandelt, Hans-Jürgen, Peter Forster, and Arne Röhl (1999). "Median-joining Networks for Inferring Intraspecific Phylogenies". In: *Molecular Biology and Evolution* 16.1, pp. 37–48 (cit. on p. 59).

Beiko, Robert G and Nicholas Hamilton (2006). "Phylogenetic Identification of Lateral Genetic Transfer Events". In: *BMC Evolutionary Biology* 6.1 (cit. on p. 126).

Bergsma, Shane and Grzegorz Kondrak (2007). "Alignment-Based Discriminative String Similarity". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (cit. on p. 111).

Betz, Werner (1949). *Deutsch und Lateinisch: Die Lehnbildungen der Althochdeutschen Benediktinerregel*. H. Bouvier (cit. on p. 4).

Bininda-Emonds, Olaf RP, John L Gittleman, and Mike A Steel (2002). "The (Super) Tree of Life: Procedures, Problems, and Prospects". In: *Annual Review of Ecology and Systematics* 33.1, pp. 265–289 (cit. on p. 91).

Boc, Alix, Anna Maria Di Sciullo, and Vladimir Makarenkov (2010a). "Classification of the Indo-European Languages Using a Phylogenetic Network Approach". In: *Classification as a Tool for Research*. Springer, pp. 647–655 (cit. on pp. 71, 72, 138).

Boc, Alix, Alpha Boubacar Diallo, and Vladimir Makarenkov (2012). "T-REX: A Web Server for Inferring, Validating and Visualizing Phylogenetic Trees and Networks". In: *Nucleic Acids Research* 40 (cit. on p. 121).

Boc, Alix, Pierre Legendre, and Vladimir Makarenkov (2013). “An Efficient Algorithm for the Detection and Classification of Horizontal Gene Transfer Events and Identification of Mosaic Genes”. In: *Algorithms from and for Nature and Life*. Springer.

Boc, Alix and Vladimir Makarenkov (2003). “New Efficient Algorithm for Detection of Horizontal Gene Transfer Events”. In: *International Workshop on Algorithms in Bioinformatics*. Springer, pp. 190–201 (cit. on pp. 72, 75, 119).

Boc, Alix, Hervé Philippe, and Vladimir Makarenkov (2010b). “Inferring and Validating Horizontal Gene Transfer Events Using Bipartition Dissimilarity”. In: *Systematic Biology* 59.2 (cit. on pp. 121–129, 133, 138, 142, 166, 170).

Boretzky, Norbert (1975). *Der Türkische Einfluss auf das Albanische. Teil 2: Wörterbuch der albanischen Turzismen*. Albanische Forschungen 12. Wiesbaden: Harrassowitz (cit. on p. 146).

Bouckaert, Remco, Joseph Heled, Denise Kühnert, et al. (2014). “BEAST 2: A Software Platform for Bayesian Evolutionary Analysis”. In: *PLoS Computational Biology* 10.4.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, et al. (2012). “Mapping the Origins and Expansion of the Indo-European Language Family”. In: *Science* 337.6097 (cit. on pp. 26, 87).

Bowern, Claire and Bethwyn Evans (2015). *The Routledge Handbook of Historical Linguistics*. Routledge.

Brodal, Gerth Stølting, Rolf Fagerberg, Thomas Mailund, Christian NS Pedersen, and Andreas Sand (2013). “Efficient Algorithms for Computing the Triplet and Quartet Distance Between Trees of Arbitrary Degree”. In: *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, pp. 1814–1832.

Brooks, Daniel R, Esra Erdem, James W Minett, and Donald Ringe (2005). “Character-based Cladistics and Answer Set Programming”. In: *International Workshop on Practical Aspects of Declarative Languages*. Springer, pp. 37–51 (cit. on p. 29).

Brown, Cecil H, Eric W Holman, Søren Wichmann, and Viveka Velupillai (2008). “Automated Classification of the World’s Languages: a Description of the Method and Preliminary Results”. In: *STUF – Language Typology and Universals* 61.4, pp. 285–308 (cit. on pp. 24, 26, 75).

Bryant, David, Flavia Filimon, and Russell D Gray (2005). “Untangling our Past: Languages, Trees, Splits and Networks”. In: *The evolution of cultural diversity: A phylogenetic approach*, pp. 67–84 (cit. on p. 58).

Bryant, David and Vincent Moulton (2004). “Neighbor-net: an Agglomerative Method for the Construction of Phylogenetic Networks”. In: *Molecular Biology and Evolution* 21.2, pp. 255–265 (cit. on p. 58).

Campbell, Lyle (2013). *Historical Linguistics: An Introduction*. 3. ed. Cambridge, Mass.: MIT Press (cit. on pp. 5, 8, 14–18, 46, 47, 60).

Cavalli-Sforza, Luigi L and Anthony WF Edwards (1967). “Phylogenetic Analysis: Models and Estimation Procedures”. In: *Evolution* 21.3, pp. 550–570.

Collins, Michael John (1996). “A New Statistical Parser Based on Bigram Lexical Dependencies”. In: *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Covington, Michael A (1998). “Alignment of Multiple Languages for Historical Comparison”. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Vol. 1 (cit. on p. 110).

Criscuolo, Alexis, Vincent Berry, Emmanuel JP Douzery, and Olivier Gascuel (2006). “SDM: A Fast Distance-Based Approach for (Super) Tree Building in Phylogenomics”. In: *Systematic Biology* 55.5, pp. 740–755 (cit. on p. 88).

Critchlow, Douglas E, Dennis K Pearl, and Chunlin Qian (1996). “The Triples Tistance for Rooted Bifurcating Phylogenetic Trees”. In: *Systematic Biology* 45.3, pp. 323–334 (cit. on p. 131).

Croft, William (2000). *Explaining Language Change: An Evolutionary Approach*. 1. publ. Longman linguistics library. Harlow: Longman (cit. on pp. 1, 22).

Dagan, Tal and William Martin (2007). “Ancestral Genome Sizes Specify the Minimum Rate of Lateral Gene Transfer During Prokaryote Evolution”. In: *Proceedings of the National Academy of Sciences* 104.3 (cit. on pp. 64, 119, 134, 170).

– (2006). “The Tree of One Percent”. In: *Genome Biology* 7.10, p. 118.

Darwin, Charles (1871). *The Descent of Man*. D. Appleton and Company (cit. on pp. 1, 6, 22).

Daval-Markussen, Aymeric and Peter Bakker (2011). “A Phylogenetic Networks Approach to the Classification of English-based Atlantic Creoles”. In: *English World-Wide* 32.2, pp. 115–146 (cit. on p. 58).

De Vaan, Michiel (2008). *Etymological Dictionary of Latin and the other Italic Languages*. Leiden, Boston, Mass. (cit. on p. 146).

Dellert, Johannes (2019a). *Information-Theoretic Causal Inference of Lexical Flow*. Vol. 4. Language Science Press (cit. on pp. 64, 120, 135–137, 171).

– (2019b). “Interactive Etymological Inference via Statistical Relational Learning”. In: *SLE 2019 Workshop Computer-assisted approaches in historical and typological language comparison*. Leipzig, Germany, p. 387 (cit. on pp. 31, 146).

Dellert, Johannes and Armin Buch (2018). “A New Approach to Concept Basicness and Stability as a Window to the Robustness of Concept List Rankings”. In: *Language Dynamics and Change* 8.2, pp. 157–181 (cit. on pp. 5, 19, 94, 115, 138, 157).

Dellert, Johannes, Thora Daneyko, Alla Münch, et al. (2020). “NorthEuraLex: a wide-coverage lexical database of Northern Eurasia”. In: *Language Resources and Evaluation* 54.1, pp. 273–301 (cit. on p. 31).

Dellert, Johannes, Gerhard Jäger, et al. (2017). *NorthEuraLex*. `http://northeuralex.org/`. Version 0.9 (cit. on pp. 20, 31, 88, 146).

Delz, Marisa (2013). “A Theoretical Approach to Automatic Loanword Detection”. PhD thesis. Eberhard-Karls-Universität Tübingen (cit. on pp. 11, 15, 26, 38, 72, 75, 121, 169).

– (2014). “Mismatches between Phylogenetic Trees in Historical Linguistics”. PhD thesis. Eberhard-Karls-Universität Tübingen (cit. on pp. 27, 63, 72, 75, 119, 131–133, 169, 170).

Delz, Marisa, Benjamin Layer, Sarah Schulz, and Johannes Wahle (2012). “Overgeneralization of Verbs – The change of the German verb system”. In: *The Evolution of Language*. World Scientific, pp. 96–103 (cit. on p. 3).

Derksen, Rick (2014). *Etymological Dictionary of the Baltic Inherited Lexicon*. Brill (cit. on p. 146).

– (2008). *Etymological Dictionary of the Slavic Inherited Lexicon*. Brill (cit. on p. 146).

Desper, Richard and Olivier Gascuel (2002). “Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle”. In: *Journal of Computational Biology* 9.5, pp. 687–705 (cit. on pp. 40, 77).

– (2004). “Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and its Relationship to Weighted Least-Squares Tree Fitting”. In: *Molecular Biology and Evolution* 21.3 (cit. on pp. 40, 77).

Doolittle, Russell F, JN Abelson, and MI Simon (1996). *Computer Methods for Macromolecular Sequence Analysis*. Vol. 266. Elsevier (cit. on p. 75).

Doolittle, W Ford (1999). “Phylogenetic Classification and the Universal Tree”. In: *Science* 284.5423, pp. 2124–2128 (cit. on p. 53).

Downey, Sean S, Brian Hallmark, Murray P Cox, Peter Norquest, and J Stephen Lansing (2008). “Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction”. In: *Journal of Quantitative Linguistics* 15.4, pp. 340–369 (cit. on pp. 80, 101).

Dudenredaktion, ed. (2013). *Etymologie der Deutschen Sprache*. 5. Auflage. Dudenverlag, Mannheim (cit. on p. 162).

Dunn, Michael, ed. (2015a). *Indo-European Lexical Cognacy Database*. `http://ielex.mpi.nl/` (cit. on p. 30).

– (2015b). “Language Phylogenies”. In: *The Routledge Handbook of Historical Linguistics*. Routledge, pp. 208–229 (cit. on pp. 23, 26, 38, 105, 106).

Dunn, Michael, Simon J Greenhill, Stephen C Levinson, and Russell D Gray (2011). “Evolved Structure of Language shows Lineage-specific Trends in Word-order Universals”. In: *Nature* 473.7345, pp. 79–82 (cit. on p. 26).

Dunning, Ted (1994). *Statistical Identification of Language*. Computing Research Laboratory, New Mexico State University (cit. on p. 111).

Dybo, Anna Vladimirovna (2013). *Ètimologičeskij slovar’ tjurkskix jazykov. Tom 9 (dopolnitel’nyj): Ètimologičeskij slovar’ bazisnoj leksiki tjurkskix jazykov*. Astana: TOO Prosper Print (cit. on p. 146).

Dyen, Isidore, Joseph B Kruskal, and Paul Black (1992). “An Indoeuropean Classification: A Lexicostatistical Experiment”. In: *Transactions of the American Philosophical Society* 82.5 (cit. on pp. 30, 84).

Edwards, Anthony WF (1964). “Reconstruction of Evolutionary Trees”. In: *Phenetic and Phylogenetic Classification* (cit. on p. 41).

Edwards, Anthony WF and Luigi L Cavalli-Sforza (1963). “The Reconstruction of Evolution”. In: *Annals of Human Genetics* 27, pp. 105–106 (cit. on p. 41).

Efron, Bradley (1979). “Bootstrap Methods: Another Look at the Jackknife”. In: *The Annals of Statistics*. Vol. 7, pp. 1–26 (cit. on p. 92).

Eldredge, Niles (2005). *Darwin: Discovering the Tree of Life*. New York: Norton (cit. on p. 1).

Embleton, Sheila M (1986). *Statistics in Historical Linguistics*. Vol. 30. Brockmeyer (cit. on pp. 19, 27, 28).

Erdem, Esra, Vladimir Lifschitz, Luay Nakhleh, and Donald Ringe (2003). “Reconstructing the Evolutionary History of Indo-European Languages Using Answer Set Programming”. In: *International Symposium on Practical Aspects of Declarative Languages*. Springer, pp. 160–176 (cit. on p. 29).

Estabrook, George F, FR McMorris, and Christopher A Meacham (1985). “Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units”. In: *Systematic Zoology* 34.2 (cit. on pp. 94, 122).

Farris, James S (1972). “Estimating Phylogenetic Trees from Distance Matrices”. In: *The American Naturalist* 106.951, pp. 645–668 (cit. on p. 141).

Felsenstein, Joseph (1978). “Cases in which Parsimony or Compatibility Methods will be Positively Misleading”. In: *Systematic Zoology* 27.4, pp. 401–410.

– (1985). “Confidence Limits on Phylogenies: An Approach using the Bootstrap”. In: *Evolution* 39.4, pp. 783–791 (cit. on p. 90).

– (1981). “Evolutionary Trees from DNA Sequences: a Maximum Likelihood Approach”. In: *Journal of Molecular Evolution* 17.6, pp. 368–376 (cit. on p. 41).

– (2004). *Inferring Phylogenies*. Vol. 2. Sunderland, MA: Sinauer Associates, Sunderland, Mass. (cit. on pp. 34, 36, 38, 40–42, 76, 89–92, 139).

– (1982). “Numerical Methods for Inferring Evolutionary Trees”. In: *The quarterly review of biology*. Vol. 57. 4. Stony Brook Foundation, Inc., pp. 379–404.

Feng, Da-Fei and Russell F Doolittle (1987). “Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees”. In: *Journal of Molecular Evolution* 25.4, pp. 351–360 (cit. on p. 109).

Fitch, Walter M (1971). “Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology”. In: *Systematic Biology* 20.4, pp. 406–416 (cit. on p. 28).

Fitch, Walter M and Emanuel Margoliash (1967). “Construction of Phylogenetic Trees”. In: *Science* 155.3760 (cit. on p. 41).

Forster, Peter, Antonio Torroni, Colin Renfrew, and Arne Röhl (2001). “Phylogenetic Star Contraction Applied to Asian and Papuan mtDNA Evolution”. In: *Molecular Biology and Evolution* 18.10, pp. 1864–1881 (cit. on p. 59).

François, Alexandre (2015). “Trees, Waves and Linkages”. In: *The Routledge Handbook of Historical Linguistics*. Routledge London, p. 161 (cit. on pp. 3, 8, 42, 86).

Gascuel, Olivier (1997). “BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data.” In: *Molecular Biology and Evolution* 14.7, pp. 685–695 (cit. on pp. 78, 104).

Gilliéron, Jules and Edmond Edmont (1902-1910). *Atlas Linguistique de la France*. Vol. 36. Champion (cit. on p. 46).

Goodman, Morris, John Czelusniak, G William Moore, Alejo E Romero-Herrera, and Genji Matsuda (1979). “Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences”. In: *Systematic Biology* 28.2, pp. 132–163 (cit. on pp. 2, 66).

Gray, Russell D and Quentin D Atkinson (2003). “Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin”. In: *Nature* 426.6965 (cit. on pp. 62, 72, 103, 105, 138).

Greenhill, Simon J, Robert Blust, and Russell D Gray (2008). “The Austronesian Basic Vocabulary Database: from Bioinformatics to Lexomics”. In: *Evolutionary Bioinformatics* 4 (cit. on p. 26).

Greenhill, Simon J, Alexei J Drummond, and Russell D Gray (2010). “How Accurate and Robust are the Phylogenetic Estimates of Austronesian Language Relationships?” In: *PLoS One* 5.3 (cit. on pp. 103, 105, 138).

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, et al. (2010). “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0”. In: *Systematic Biology* 59.3, pp. 307–321 (cit. on p. 104).

Haeckel, Ernst H. P. A. (1874). *Anthropogenie oder Entwicklungsgeschichte des Menschen*. Leipzig: Verlag von Wilhelm Engelmann.

Hallett, Michael T and Jens Lagergren (2001). “Efficient Algorithms for Lateral Gene Transfer Problems”. In: *Proceedings of the Fifth Annual International Conference on Computational Biology*, pp. 149–156 (cit. on pp. 75, 126).

Hamed, Mahe Ben and Frend Wang (2006). “Stuck in the Forest: Trees, Networks and Chinese Dialects”. In: *Diachronica* 23.1, pp. 29–60 (cit. on p. 58).

Hammarström, Harald, Sebastian Bank, Robert Forkel, and Martin Haspelmath (2018). *Glottolog 3.2*. `http://glottolog.org/`. Accessed: 2018-02-18. Max Planck Institute for the Science of Human History, Jena (cit. on pp. 45, 88, 94, 137).

Haspelmath, Martin (2009). “Lexical Borrowing: Concepts and Issues”. In: *Loanwords in the World’s Languages: A Comparative Handbook*, pp. 35–54 (cit. on pp. 14, 16, 18, 19).

– (2008). “Loanword Typology: Steps toward a Systematic Cross-Linguistic Study of Lexical Borrowability”. In: *Empirical Approaches to Language Typology* 35, p. 43 (cit. on pp. 11, 14).

Haspelmath, Martin and Uri Tadmor (2009a). “The Loanword Typology Project and the World Loanword Database”. In: *Loanwords in the World’s Languages: A Comparative Handbook*, pp. 1–34 (cit. on p. 32).

– eds. (2009b). *WOLD*. `https://wold.clld.org/`. Leipzig: Max Planck Institute for Evolutionary Anthropology (cit. on pp. 19, 32, 146, 154, 157, 160).

Haugen, Einar (1950). “The Analysis of Linguistic Borrowing”. In: *Language* 26.2, pp. 210–231 (cit. on pp. 4, 12–14, 19, 47, 71, 76, 167).

Heggarty, Paul, Warren Maguire, and April McMahon (2010). “Splits or Waves? Trees or Webs? How Divergence Measures and Network Analysis can unravel Language Histories”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1559, pp. 3829–3843 (cit. on pp. 58, 59).

Heled, Joseph and Remco R Bouckaert (2013). “Looking for Trees in the Forest: Summary Tree from Posterior Samples”. In: *BMC Evolutionary Biology* 13.1 (cit. on pp. 99, 100, 138).

Hennig, Willi (1999). *Phylogenetic Systematics*. University of Illinois Press (cit. on p. 22).

Hickey, Raymond (2012). *The Handbook of Language Contact*. John Wiley & Sons (cit. on p. 9).

Hinneburg, Alexander, Heikki Mannila, Samuli Kaislaniemi, Terttu Nevalainen, and Helena Raumolin-Brunberg (2007). “How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change”. In: *Literary and Linguistic Computing* 22.2, pp. 137–150 (cit. on p. 103).

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Sy Vinh Le (2017). “UFBoot2: Improving the Ultrafast Bootstrap Approximation”. In: *MMolecular Biology and Evolution* (cit. on p. 115).

Hogeweg, Paulien and Ben Hesper (1984). “The Alignment of Sets of Sequences and the Construction of Phyletic Trees: An Integrated Method”. In: *Journal of Molecular Evolution* 20.2, pp. 175–186 (cit. on p. 109).

Holman, Eric W, Søren Wichmann, Cecil H Brown, et al. (2008). “Advances in Automated Language Classification”. In: *Quantitative Investigations in Theoretical Linguistics* (cit. on pp. 5, 7, 19, 26, 31, 75, 78).

Huelsenbeck, John P and Fredrik Ronquist (2001). “MrBayes: Bayesian Inference of Phylogeny”. In: *Bioinformatics* 17 (cit. on p. 106).

Huelsenbeck, John P, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback (2001). “Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology”. In: *Science* 294.5550 (cit. on p. 105).

Huffman, Stephen M (1998). *The Genetic Classification of Languages by N-Gram Analysis: A Computational Technique*. Georgetown University (cit. on p. 111).

Huson, Daniel H and David Bryant (2005). “Application of Phylogenetic Networks in Evolutionary Studies”. In: *Molecular Biology and Evolution* 23.2, pp. 254–267.

– (2006). “Application of Phylogenetic Networks in Evolutionary Studies”. In: *Molecular Biology and Evolution* 23.2, pp. 254–267 (cit. on p. 58).

Huson, Daniel H, Regula Rupp, and Celine Scornavacca (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press (cit. on pp. 35–42, 44, 53–56, 58, 62, 76, 77, 104, 106, 122, 140).

Huson, Daniel H and Celine Scornavacca (2012). “Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks”. In: *Systematic Biology* 61.6, pp. 1061–1067 (cit. on p. 63).

International Phonetic Association, ed. (2019). *IPA Chart*. `https://www.internationalphoneticassociation.org/content/full-ipa-chart`. Version 2019. Accessed: 2020-03-18 (cit. on p. 23).

Itkonen, Erkki and Ulla-Maija Kulonen (1992). *Suomen sanojen alkuperä. Etymologinen sanakirja 1-3*. Vol. 2012 (cit. on p. 146).

Jäger, Gerhard (2019). “Computational Historical Linguistics”. In: *Theoretical Linguistics* 45 (cit. on p. 23).

– (2013a). “Evaluating Distance-Based Pyhlogenetic Algorithms for Automated Language Classification”. In: *Manuscript, University of Tübingen* (cit. on pp. 7, 26, 44, 45, 75, 77, 78, 88, 94, 95, 137, 193).

Jäger, Gerhard (2018). “Global-scale Phylogenetic Linguistic Inference from Lexical Resources”. In: *Scientific Data* 5.1, pp. 1–16.

Jäger, Gerhard (2013b). “Phylogenetic Inference from Word Lists using Weighted Alignment with Empirically Determined Weights”. In: *Language Dynamics and Change* 3.2, pp. 245–291 (cit. on pp. 7, 26, 45, 75, 78–80, 85, 94, 95, 102, 110, 137).

Jäger, Gerhard (2019). *Rooting MADness*. Computer-Assisted Language Comparison in Practice. `https://calc.hypotheses.org/1899` Accessed: 2020-07-24 (cit. on pp. 140, 142).

Jäger, Gerhard and Johann-Mattis List (2015). *Factoring Lexical and Phonetic Phylogenetic Characters from Word Lists*. Universitätsbibliothek Tübingen (cit. on pp. 108–110, 117).

– (2016). *Statistical and Computational Elaborations of the Classical Comparative Method*. `http://www.sfs.uni-tuebingen.de/~gjaeger/publications/jaegerListOxfordHandbook.pdf` (cit. on pp. 5, 22, 23, 25, 26, 38, 76, 106, 109, 110).

– (2018). “Using Ancestral State Reconstruction Methods for Onomasiological Reconstruction in Multilingual Word Lists”. In: *Language Dynamics and Change* 8.1, pp. 22–54 (cit. on pp. 64, 134, 170).

Jäger, Gerhard and Søren Wichmann (2016). “Inferring the World Tree of Languages from Word Lists”. In: *The Evolution of Language: Proceedings of the 11th International Conference on the Evolution of Language (EvoLang XI), ed. by Séan G. Roberts, Christine Cuskley, Luke McCrohon, Lluis Barceló-Coblijn, Olga Feher & Tessa Verhoef. Available at evolang.org/neworleans/papers/147.html* (cit. on p. 113).

Janhunen, Juha (1977). *Samojedischer Wortschatz: Gemeinsamojedische Etymologien*. Castrenianumin Toimitteita. Helsinki (cit. on p. 146).

Janhunen, Juha et al. (1981). *Uralilaisen kantakielen sanastosta*. Suomalais-ugrilainen seura (cit. on p. 146).

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler, and Lars S Jermiin (2017). “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates”. In: *Nature Methods* 14.6, p. 587 (cit. on p. 104).

Kessler, Brett (2001). *The Significance of Word Lists*. Center for the Study of Language and Information (cit. on pp. 13, 19, 51, 58, 71, 165).

Kleiweg, Peter, John Nerbonne, and Leonie Bosveld (2004). “Geographic Projection of Cluster Composites”. In: *International Conference on Theory and Application of Diagrams*. Springer.

Koellner, Marisa and Johannes Dellert (2016). “Ancestral State Reconstruction and Loanword Detection”. In: *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Universität Tübingen (cit. on p. 133).

– (2017). “Topological Tree Comparison and Ancestral State Reconstruction for Loanword Detection”. Unpublished (cit. on p. 133).

Kroonen, Guus (2013). *Etymological Dictionary of Proto-Germanic*. Brill (cit. on pp. 146, 154, 157, 160).

Kundu, Soumya and Mukul S Bansal (2018). “On the Impact of Uncertain Gene Tree Rooting on Duplication-Transfer-Loss Reconciliation”. In: *BMC Bioinformatics* 19.9, pp. 21–31.

Lapointe, François-Joseph, John AW Kirsch, and Robert Bleiweiss (1994). “Jackknifing of Weighted Trees: Validation of Phylogenies Reconstructed from Distance Matrices”. In: *Molecular Phylogenetics and Evolution* 3.3, pp. 256–267.

Lecointre, Guillaume (2006). *The Tree of Life: a Phylogenetic Classification*. Ed. by Hervé Le Guyader. Cambridge, Mass.: Belknap Press of Harvard University Press (cit. on p. 22).

Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, et al. (2014). “Behind Family Trees: Secondary Connections in Uralic Language Networks”. In: *Language Dynamics and Change* 4.2, pp. 189–221 (cit. on pp. 58, 59).

Lehtiranta, Juhani (2001). *Yhteissaamelainen sanasto*. Suomalais-ugrilaisen Seuran toimituksia. Helsinki: Suomalais-Ugrilainen Seura (cit. on pp. 146, 154).

Lemey, Philippe, Marco Salemi, and Anne-Mieke Vandamme (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.

Levenshtein, Vladimir I (1966). “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”. In: *Soviet physics doklady*. Vol. 10. 8.

List, Johann-Mattis (2019). “Automated Methods for the Investigation of Language Contact, with a Focus on Lexical Borrowing”. In: *Language and Linguistics Compass* 13.10 (cit. on pp. 28, 76, 145, 167).

– (2012). “Multiple Sequence Alignment in Historical Linguistics”. In: *Proceedings of ConSOLE*. Vol. 19, pp. 241–260 (cit. on p. 110).

– (2015). “Network Perspectives on Chinese Dialect History: Chances and Challenges”. In: *Bulletin of Chinese Linguistics* 8.1, pp. 27–47 (cit. on pp. 64, 134, 170).

– (2010). “SCA: Phonetic Alignment Based on Sound Classes”. In: *New Directions in Logic, Language and Computation*. Springer, pp. 32–51 (cit. on p. 110).

– (2014). *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press (cit. on p. 26).

List, Johann-Mattis, Simon Greenhill, and Robert Forkel (2017). *LingPy. A Python Library for Quantitative Tasks in Historical Linguistics*. `http://lingpy.org` (cit. on p. 134).

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler, and William Martin (2014a). “Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution”. In: *Bioessays* 36.2, pp. 141–150 (cit. on pp. 64, 134, 135, 170).

List, Johann-Mattis and Nathanael E Schweikhard (2020). “Modeling Word Trees in Historical Linguistics: Preliminary Ideas for the Reconciliation of Word Trees and Language Trees”. Unrevised preprint, submitted for the Sektionsband Historische Linguistik (cit. on pp. 50, 51).

List, Johann-Mattis, Nelson-Sathi Shijulal, William Martin, and Hans Geisler (2014b). “Using Phylogenetic Networks to Model Chinese Dialect History”. In: *Language Dynamics and Change* 4.2 (cit. on pp. 64, 134, 170).

Liu, Liang and Dennis K Pearl (2007). “Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions”. In: *Systematic Biology* 56.3, pp. 504–514 (cit. on p. 75).

MacLeod, Dave, Robert L Charlebois, Ford Doolittle, and Eric Bapteste (2005). “Deduction of Probable Events of Lateral Gene Transfer Through Comparison of Phylogenetic Trees by Recursive Consolidation and Rearrangement”. In: *BMC Evolutionary Biology* 5.1, pp. 1–11 (cit. on p. 126).

Maddison, Wayne P (1997). “Gene Trees in Species Trees”. In: *Systematic Biology* 46.3, pp. 523–536 (cit. on pp. 66, 129).

Makarenkov, Vladimir, Alix Boc, Charles F Delwiche, Hervé Philippe, et al. (2006). “New Efficient Algorithm for Modeling Partial and Complete Gene Transfer Scenarios”. In: *Data Science and Classification*. Springer (cit. on pp. 72, 75, 120, 121, 126).

Margush, Timothy and Fred R McMorris (1981). “Consensus-Trees”. In: *Bulletin of Mathematical Biology* 43.2, pp. 239–244.

McMahon, April (2010). “Computational Models and Language Contact”. In: *The Handbook of Language Contact*. Wiley Online Library, p. 128 (cit. on pp. 19, 46).

McMahon, April, Paul Heggarty, Robert McMahon, and Warren Maguire (2007). “The Sound Patterns of English: Representing Phonetic Similarity”. In: *English Language & Linguistics* 11.1, pp. 113–142 (cit. on p. 58).

McMahon, April, Paul Heggarty, Robert McMahon, and Natalia Slaska (2005). “Swadesh Sublists and the Benefits of Borrowing: an Andean Case Study 1”. In: *Transactions of the Philological Society* 103.2, pp. 147–170 (cit. on p. 59).

McMahon, April and Robert McMahon (2005). *Language Classification by Numbers*. Oxford University Press on Demand (cit. on pp. 28, 58–60, 97, 130).

Mennecier, Philippe, John Nerbonne, Evelyne Heyer, and Franz Manni (2016). “A Central Asian Language Survey: Collecting Data, Measuring Relatedness and Detecting Loans”. In: *Language Dynamics and Change* 6.1, pp. 57–98 (cit. on p. 29).

Merriam-Webster Dictionary (2020). *mountain*. `https://www.merriam-webster.com/`. Accessed: 2020-03-25 (cit. on pp. 20, 50, 70, 157).

Meyer-Lübke, Wilhelm (2009). *Romanisches Etymologisches Wörterbuch*. Vol. 7. Sammlung romanischer Elementar- und Handbücher, 3. Reihe, Band 3. Winter, Heidelberg (cit. on p. 146).

Minett, James W and William SY Wang (2003). “On Detecting Borrowing: Distance-based and Character-based Approaches”. In: *Diachronica* 20.2, pp. 289–330 (cit. on p. 28).

Minh, Bui Quang, Minh Anh Thi Nguyen, and Arndt von Haeseler (2013). “Ultrafast Approximation for Phylogenetic Bootstrap”. In: *Molecular Biology and Evolution* 30.5, pp. 1188–1195.

Mirkin, Boris, Ilya Muchnik, and Temple F Smith (1995). “A Biologically Consistent Model for Comparing Molecular Phylogenies”. In: *Journal of Computational Biology* 2.4, pp. 493–507 (cit. on p. 66).

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*. MIT press.

Moore, G William, M Goodman, and J Barnabas (1973). “An Iterative Approach from the Standpoint of the Additive Hypothesis to the Dendrogram Problem Posed by Molecular Data Sets”. In: *Journal of Theoretical Biology* 38.3, pp. 423–457 (cit. on p. 40).

Moret, Bernard ME, Luay Nakhleh, Tandy Warnow, et al. (2004). “Phylogenetic Networks: Modeling, Reconstructibility, and Accuracy”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1.1, pp. 13–23.

Morrison, David A (2011). *Introduction to Phylogenetic Networks*. Uppsala, Sweden: RJR Productions (cit. on pp. 33, 36, 54–56, 60, 63–68, 71, 167).

Mueller, Laurence D and Francisco J Ayala (1982). “Estimation and Interpretation of Genetic Distance in Empirical Studies”. In: *Genetics Research* 40.2, pp. 127–137 (cit. on p. 90).

Muysken, Pieter C (1981). “Halfway between Quechua and Spanish: The case for Relexification”. In: *Historicity and variation in creole studies*.

Myers-Scotton, Carol et al. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford University Press on Demand (cit. on p. 12).

Nakhleh, Luay, Don Ringe, and Tandy Warnow (2005a). “Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages”. In: *Language*, pp. 382–420 (cit. on pp. 29, 126, 127).

Nakhleh, Luay, Derek Ruths, and Hideki Innan (2009). “GeneTrees, Species Trees, and Species Networks”. In: *Citeseer* (cit. on p. 75).

Nakhleh, Luay, Derek Ruths, and Li-San Wang (2005b). “RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer”. In: *International Computing and Combinatorics Conference*. Springer, pp. 84–93 (cit. on p. 75).

Needleman, Saul B and Christian D Wunsch (1970). “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins”. In: *Journal of Molecular Biology* 48.3 (cit. on p. 78).

Nei, Masatoshi and Sudhir Kumar (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, et al. (2011). “Networks Uncover Hidden Lexical Borrowing in Indo-European Language Evolution”. In: *Proceedings of the Royal Society B: Biological Sciences* 278.1713, pp. 1794–1803 (cit. on pp. 64, 119, 134, 170).

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni (2008). “Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering”. In: *Data Analysis, Machine Learning and Applications* (cit. on p. 93).

Newman, Mark EJ and Michelle Girvan (2004). “Finding and Evaluating Community Structure in Networks”. In: *Physical Review E* 69.2 (cit. on p. 84).

Nguyen, Lam-Tung, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh (2014). “IQ-TREE: A Fast and Effective Stochastic Algorithm for estImating Maximum-Likelihood Phylogenies”. In: *Molecular Biology and Evolution* 32.1, pp. 268–274 (cit. on pp. 104, 115).

Nichols, Johanna and Tandy Warnow (2008). “Tutorial on Computational Linguistic Phylogeny”. In: *Language and Linguistics Compass* 2.5, pp. 760–820.

Noonan, Michael (2010). “Genetic Classification and Language Contact”. In: *The Handbook of Language Contact*. Wiley Online Library, pp. 48–65 (cit. on pp. 6, 7, 26, 42).

Notredame, Cédric, Desmond G Higgins, and Jaap Heringa (2000). “T-Coffee: a Novel Method for Fast and Accurate Multiple Sequence Alignment”. In: *Journal of Molecular Biology* 302.1, pp. 205–217 (cit. on p. 109).

O’Grady, William, John Archibad, and Francis Katamba (2011). *Contemporary Linguistics: An Introduction*. Vol. 2. Longman (cit. on p. 23).

Orel, Vladimir (1998). *Albanian Etymological Dictionary*. Brill (cit. on p. 146).

Page, Roderic DM (1994a). "Maps Between Trees and Cladistic Analysis of Historical Associations among Genes, Organisms, and Areas". In: *Systematic Biology* 43.1, pp. 58–77 (cit. on p. 66).

– (1993). "On Islands of Trees and the Efficacy of Different Methods of Branch Swapping in Finding Most-Parsimonious Trees". In: *Systematic Biology*, pp. 200–210 (cit. on p. 40).

– (1994b). "Parallel Phylogenies: Reconstructing the History of Host-Parasite Assemblages". In: *Cladistics* 10.2, pp. 155–173 (cit. on p. 66).

Page, Roderic DM and Michael A Charleston (1997). "From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem". In: *Molecular Phylogenetics and Evolution* 7.2, pp. 231–240 (cit. on p. 66).

– (1998). "Trees Within Trees: Phylogeny and Historical Associations". In: *Trends in Ecology & Evolution* 13.9, pp. 356–359 (cit. on p. 129).

Pamilo, Pekka and Masatoshi Nei (1988). "Relationships between Gene Trees and Species Trees". In: *Molecular Biology and Evolution* 5.5, pp. 568–583.

Paradis, Carole and Darlene LaCharité (1997). "Preservation and Minimality in Loanword Adaptation". In: *Journal of Linguistics* 33.02, pp. 379–430 (cit. on p. 15).

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R Language". In: *Bioinformatics* 20, pp. 289–290 (cit. on pp. 78, 88).

Paul, Hermann (1886). *Principien der Sprachgeschichte*. Niemeyer (cit. on p. 4).

Penny, David and MD Hendy (1986). "Estimating the Reliability of Evolutionary Trees." In: *Molecular Biology and Evolution* 3.5, pp. 403–417.

– (1985). "Testing Methods of Evolutionary Tree Construction". In: *Cladistics* 1.3, pp. 266–278.

Polzin, Tobias and Siavash Vahdati Daneshmand (2003). "On Steiner Trees and Minimum Spanning Trees in Hypergraphs". In: *Operations Research Letters* 31.1, pp. 12–20 (cit. on p. 59).

Pompei, Simone, Vittorio Loreto, and Francesca Tria (2011). "On the Accuracy of Language Trees". In: *PLoS One* 6.

Prokic, Jelena (2010). *Families and Resemblances*. Ph. D. thesis, Rijksuniversiteit Groningen (cit. on p. 58).

Prokić, Jelena, Martijn Wieling, and John Nerbonne (2009). "Multiple Sequence Alignments in Linguistics". In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 18–25 (cit. on p. 110).

Rama, Taraka (2015a). "Automatic Cognate Identification with Gap-Weighted String Subsequences." In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rama, Taraka (2015b). “Studies in Computational Historical Linguistics: Models and Analyses”. PhD thesis. Göteborgs universitet. Humanistiska fakulteten.

Rama, Taraka and Lars Borin (2014). “N-Gram Approaches to the Historical Dynamics of Basic Vocabulary”. In: *Journal of Quantitative Linguistics* 21.1 (cit. on p. 111).

Rama, Taraka, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger (2018). “Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics?” In: *arXiv preprint arXiv:1804.05416* (cit. on pp. 102, 116, 162).

Rama, Taraka, Johannes Wahle, Pavel Sofroniev, and Gerhard Jäger (2017). “Fast and Unsupervised Methods for Multilingual Cognate Clustering”. In: *arXiv preprint arXiv:1702.04938* (cit. on pp. 26, 79, 80, 84, 86, 101, 106, 138, 143, 144).

Rasmussen, Matthew D and Manolis Kellis (2007). “Accurate Gene-Tree Reconstruction by Learning Gene-and Species-Specific Substitution Rates Across Multiple Complete Genomes”. In: *Genome Research* 17.12, pp. 1932–1942 (cit. on p. 75).

Ringe, Don, Tandy Warnow, and Ann Taylor (2002). “Indo-European and Computational Cladistics”. In: *Transactions of the Philological Society* 100.1, pp. 59–129 (cit. on pp. 26, 29, 30).

Ringe, Don, Tandy Warnow, Ann Taylor, Alexander Michailov, and Libby Levison (1998). “Computational Cladistics and the Position of Tocharian”. In: *The Bronze Age and early Iron Age peoples of Eastern Central Asia* 1, pp. 391–414.

Robinson, David F (1971). “Comparison of Labeled Trees with Valency Three”. In: *Journal of Combinatorial Theory, Series B* 11.2, pp. 105–119 (cit. on p. 40).

Robinson, David F and Leslie R Foulds (1981). “Comparison of Phylogenetic Trees”. In: *Mathematical Biosciences* 53 (cit. on pp. 122, 131).

Ronquist, Fredrik and John P Huelsenbeck (2003). “MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models”. In: *Bioinformatics* 19.12 (cit. on p. 106).

Rose, Yvan (2012). “Perception, Representation, and Correspondence Relations in Loanword Phonology”. In: *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*. Vol. 25. 1 (cit. on p. 15).

Saitou, Naruya and Masatoshi Nei (1987). “The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.” In: *Molecular Biology and Evolution* 4, pp. 406–425 (cit. on p. 39).

Salemi, Marco, Philippe Lemey, and Anne-Mieke Vandamme (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press (cit. on p. 41).

Sammallahti, Pekka (1988). “The Uralic Languages: Description, History and Foreign Influences”. In: *The Uralic Languages: Description, History and Foreign Influences. Leiden*, pp. 478–554 (cit. on p. 146).

Sand, Andreas, Morten K Holt, Jens Johansen, et al. (2013). “Algorithms for Computing the Triplet and Quartet Distances for Binary and General Trees”. In: *Biology* 2.4, pp. 1189–1209 (cit. on p. 132).

Sand, Andreas, Morten K Holt, Jens Johansen, et al. (2014). “tqDist: A Library for Computing the Quartet and Triplet Distances Between Binary or General Trees”. In: *Bioinformatics* 30.14.

Sankoff, David (1975). “Minimal Mutation Trees of Sequences”. In: *SIAM Journal on Applied Mathematics* 28.1, pp. 35–42 (cit. on p. 41).

Sankoff, David and Robert J Cedergren (1983). “Simultaneous Comparison of Three or More Sequences Related by a Tree”. In: *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*.

Sankoff, David, Cristiane Morel, and Robert J Cedergren (1973). “Evolution of 5S RNA and the Non-Randomness of Base Replacement”. In: *Nature New Biology* 245.147.

Sanžeev, Garma Dancaranovič, Marija Nikolaevna Orlovskaja, and Zoja Vasil’evna Ševerina (2015). *Ètimologičeskij slovar’ mongol’skix jazykov: v 3-x tomax*. 3 vols. Moscow: Institut vostokovedenija RAN (cit. on p. 146).

Schleicher, August (1861). *Compendium der vergleichenden Grammatik der Indogermanischen Sprachen*. Böhlau (cit. on p. 1).

– (1863). *Die Darwinsche Theorie und die Sprachwissenschaft*. Weimar: Hermann Böhlau (cit. on pp. 1, 4, 6, 22).

Schlie, Klaus P (2011). “Phangorn: Phylogenetic Analysis in R”. In: *Bioinformatics* 27.4, pp. 592–593.

Schliep, Klaus P, Alastair J Potts, David A Morrison, and Guido W Grimm (2017). “Intertwining Phylogenetic Trees and Networks”. In: *Methods in Ecology and Evolution* 8.10, pp. 1212–1220 (cit. on p. 58).

Schmidt, Johannes (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. H. Böhlau (cit. on p. 8).

Schuchardt, Hugo (1868). *Vokalismus*. Teubner Leipzig (cit. on pp. 4, 8, 61).

Seiler, Friedrich (1907-1913). *Die Entwicklung der Deutschen Kultur im Spiegel des Deutschen Lehnworts*. Vol. 4. Buchhandlung des Waisenhauses (cit. on p. 4).

Sevortyan, EV (1974). “Etymological Dictionary of Turkic Languages”. In: *M.: Nauka* 1992 (cit. on p. 146).

Silverman, Daniel (1992). “Multiple Scansions in Loanword Phonology: Evidence from Cantonese”. In: *Phonology* 9.2, pp. 298–328 (cit. on p. 15).

Singh, Anil Kumar and Harshit Surana (2007). “Can Corpus Based Measures be Used for Comparative Study of Languages?” In: *Proceedings of 9th Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*. Association for Computational Linguistics (cit. on p. 111).

Singh, Rajendra (1981). “Aspects of Language Borrowing: English Loans in Hindi”. In: *Sprachkontakt und Sprachkonflikt*, pp. 113–116 (cit. on p. 12).

Smits, Caroline (1998). “Two Models for the Study of Language Contact”. In: *Historical Linguistics, 1997: Selected Papers from the 13th International Conference on Historical Linguistics, Düsseldorf, 10-17 August 1997*. Vol. 164. John Benjamins Publishing, p. 377 (cit. on p. 10).

Sokal, Robert R (1958). “A Statistical Method for Evaluating Systematic Relationship”. In: *University of Kansas Science Bulletin* 28, pp. 1409–1438 (cit. on p. 39).

Stamatakis, Alexandros (2014). “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies”. In: *Bioinformatics* 30.9 (cit. on p. 104).

Steel, Mike (2016). *Phylogeny: Discrete and Random Processes in Evolution*. SIAM (cit. on p. 37).

Steiner, Lydia, Michael Cysouw, and Peter Stadler (2011). “A Pipeline for Computational Historical Linguistics”. In: *Language Dynamics and Change* 1.1, pp. 89–127 (cit. on pp. 45, 95, 137).

Sukumaran, Jeet and Mark T Holder (2010). “DendroPy: a Python Library for Phylogenetic Computing”. In: *Bioinformatics* 26.12, pp. 1569–1571 (cit. on p. 100).

Swadesh, Morris (1971). *The Origin and Diversification of Language*. Transaction Publishers.

– (1955). “Towards greater Accuracy in Lexicostatistic Dating”. In: *International Journal of American Linguistics* 21.2, pp. 121–137 (cit. on pp. 5, 19).

Szeto, Pui Yiu, Umberto Ansaldo, and Stephen Matthews (2018). “Typological Variation across Mandarin Dialects: An Areal Perspective with a Quantitative Approach”. In: *Linguistic Typology* 22.2, pp. 233–275 (cit. on p. 58).

Tadmor, Uri, Martin Haspelmath, and Bradley Taylor (2010). “Borrowability and the Notion of Basic Vocabulary”. In: *Diachronica* 27.2, pp. 226–246 (cit. on pp. 11, 12, 19, 172).

Than, Cuong, Guohua Jin, and Luay Nakhleh (2008a). “Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer”. In: *RECOMB International Workshop on Comparative Genomics*. Springer, pp. 113–127.

Than, Cuong and Luay Nakhleh (2008). “SPR-Based Tree Reconciliation: Non-Binary Trees and Multiple Solutions”. In: *Proceedings of the 6th Asia-Pacific Bioinformatics Conference*. World Scientific, pp. 251–260 (cit. on p. 127).

Than, Cuong, Derek Ruths, Hideki Innan, and Luay Nakhleh (2007). “Confounding Factors in HGT Detection: Statistical Error, Coalescent Effects, and Multiple Solutions”. In: *Journal of Computational Biology* 14.4, pp. 517–535 (cit. on p. 126).

Than, Cuong, Derek Ruths, and Luay Nakhleh (2008b). “PhyloNet: A Software Package for Analyzing and Reconstructing Reticulate Evolutionary Relationships”. In: *BMC Bioinformatics* 9.1 (cit. on pp. 127, 128).

Thomason, Sarah G (2010). “Contact Explanations in Linguistics”. In: *The Handbook of Language Contact*. Wiley Online Library, pp. 31–47 (cit. on p. 9).

Thomason, Sarah G and Terrence Kaufman (1992). *Language Contact, Creolization, and Genetic Linguistics*. University of California Press (cit. on pp. 9–11).

Tofigh, Ali (2009). “Using Trees to Capture Reticulate Evolution”. PhD thesis. KTH School of Computer Science and Communication.

Tofigh, Ali, Michael Hallett, and Jens Lagergren (2010). “Simultaneous Identification of Duplications and Lateral Gene Transfers”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.2, pp. 517–535.

Tria, Fernando Domingues Kümmel, Giddy Landan, and Tal Dagan (2017). “Phylogenetic Rooting Using Minimal Ancestor Deviation”. In: *Nature Ecology & Evolution* 1.1, pp. 1–7 (cit. on pp. 140, 142).

Van Coetsem, Frans (2000). *A General and Unified Theory of the Transmission Process in Language Contact*. Vol. 19. Winter (cit. on p. 9).

– (1988). *Loan Phonology and the Two Transfer Types in Language Contact*. Foris (cit. on pp. 9, 10).

Van Der Ark, René, Philippe Mennecier, John Nerbonne, and Franz Manni (2007). “Preliminary Identification of Language Groups and Loan Words in Central Asia”. In: *Proceedings of the RANLP Workshop on Computational Phonology. Borovetz: RANLP*, pp. 13–20 (cit. on pp. 29, 76).

Van Hout, Roeland and Pieter Muysken (1994). “Modeling Lexical Borrowability”. In: *Language Variation and Change* 6.1, pp. 39–62 (cit. on p. 11).

Vandamme, Anne-Mieke (2009). “Basic Concepts of Molecular Evolution”. In: *The Phylogenetic Handbook* 2, pp. 3–32 (cit. on p. 141).

Vendelin, Inga and Sharon Peperkamp (2004). “Evidence for Phonetic Adaptation of Loanwords: An Experimental Study”. In: *Actes des Journées d’Etudes Linguistique*, pp. 129–131 (cit. on p. 15).

Volland, Brigitte (1986). *Französische Entlehnungen im Deutschen: Transferenz und Integration auf phonologischer, graphematischer, morphologischer und lexikalisch-semantischer Ebene*. Vol. 163. Walter de Gruyter (cit. on p. 14).

Wang, William SY and James W Minett (2005). “Vertical and Horizontal Transmission in Language Evolution”. In: *Transactions of the Philological Society* 103.2, pp. 121–146 (cit. on pp. 28, 29).

Weinreich, Uriel (1968). *Languages in Contact: Findings and Problems*. 6th edition. The Hague: Mouton (cit. on p. 11).

Weiss, Michael (2015). “The Comparative Method”. In: *The Routledge Handbook of Historical Linguistics*. Routledge, pp. 145–163 (cit. on pp. 4, 5).

Whitney, William D (1881). “On Mixture in Language”. In: *Transactions of the American Philological Association (1869–1896)* 12, pp. 5–26 (cit. on p. 11).

Wichman, Søren, Eric W Holman, and Cecil H Brown (2018). *The ASJP Database*. `http://asjp.clld.org/`. Version 18 (cit. on pp. 24, 31, 87, 88).

Wichmann, Søren, André Müller, and Viveka Velupillai (2010). “Homelands of the World’s Language Families: A Quantitative Approach”. In: *Diachronica* 27.2, pp. 247–276 (cit. on p. 87).

Wichmann, Søren, Taraka Rama, and Eric W Holman (2011). “Phonological Diversity, Word Length, and Population Sizes Across Languages: The ASJP Evidence”. In: *Linguistic Typology* 15.2, pp. 177–197 (cit. on p. 87).

Willems, Matthieu, Etienne Lord, Louise Laforest, et al. (2016). “Using Hybridization Networks to Retrace the Evolution of Indo-European Languages”. In: *BMC evolutionary biology* 16.1, p. 180 (cit. on pp. 60–63, 75, 82, 84–86, 100, 169).

Willems, Matthieu, Nadia Tahiri, and Vladimir Makarenkov (2014). “A New Efficient Algorithm for Inferring Explicit Hybridization Networks Following the Neighbor-Joining Principle”. In: *Journal of Bioinformatics and Computational Biology* 12.05, p. 1450024 (cit. on p. 62).

Winford, Donald (2010). “Contact and Borrowing”. In: *The Handbook of Language Contact*. Vol. 170. Wiley Online Library, p. 187 (cit. on pp. 10–12, 14).

– (2005). “Contact-induced Changes: Classification and Processes”. In: *Diachronica* 22.2, pp. 373–427 (cit. on p. 10).

Wodtko, Dagmar S, Britta Irslinger, and Carolin Schneider (2008). *Nomina im Indogermanischen Lexikon*. Heidelberg: Winter (cit. on p. 50).

Wu, Chien-Fu Jeff (1986). “Jackknife, Bootstrap and other Resampling Methods in Regression Analysis”. In: *The Annals of Statistics*, pp. 1261–1295.

Yallop, Colin and Janet Fletcher (2007). *An Introduction to Phonetics and Phonology*. 3. ed. Blackwell Publishing (cit. on p. 23).

Zaicz, Gábor (2006). *Etimológiai szótár: Magyar szavak és toldalékok eredete*. Tinta Könyvk. (cit. on p. 146).

Appendix A

A.1 Evidence for Noisy Bootstrap: Neighbor-Joining versus FastME

All analyses in the experiment to verify the noisy bootstrap method presented in chapter 4.1.4 are also carried out using the neighbor-joining algorithm. The comparison of the results is shown in table A.1.

Inferred language tree	GQD
NJ	0.0324
FastMe	**0.0193**
Consensus tree standard bootstrap	GQD
NJ	0.0250
FastMe	**0.0193**
Consensus tree noisy bootstrap	GQD
NJ	0.0317
FastMe	**0.0206**

Tab. A.1.: Generalized quartet distance between the expert tree and inferred trees obtained from the NJ algorithm.

The generalized quartet distances (GQD) show a high agreement with the Glottolog tree for both methods. Using NJ, the agreement between the automatically reconstructed NJ tree is slightly smaller, but still indicates a correctness of around 80%. A similar distance is computed between the consensus tree obtained from the noisy bootstrap method and the expert tree. The results underline the usage of the noisy bootstrap method as suitable alternative to the standard bootstrap method in terms of computing the accuracy of trees.
The comparison of the reconstruction methods shows a higher agreement between the trees inferred using the FastME algorithm. This underlines the results of Jäger (2013a), who showed that the FastME algorithm with post-processing options slightly improves the results of automatic language classification compared to other reconstruction methods. This emphasizes FastME as the method of choice for the reconstruction of distance-based trees.

A.2 Automatically Inferred Language Tree on NELex

The following figure represents the automatically reconstructed language tree for the NELex database using MrBayes. The tree is the best representative tree from the posterior distribution, i.e. the MCC tree. It is an actual tree from the sample with a fully resolved topology. In addition, the tree is rooted using the MAD rooting algorithm. For a better representation, the tree is split into two halves.

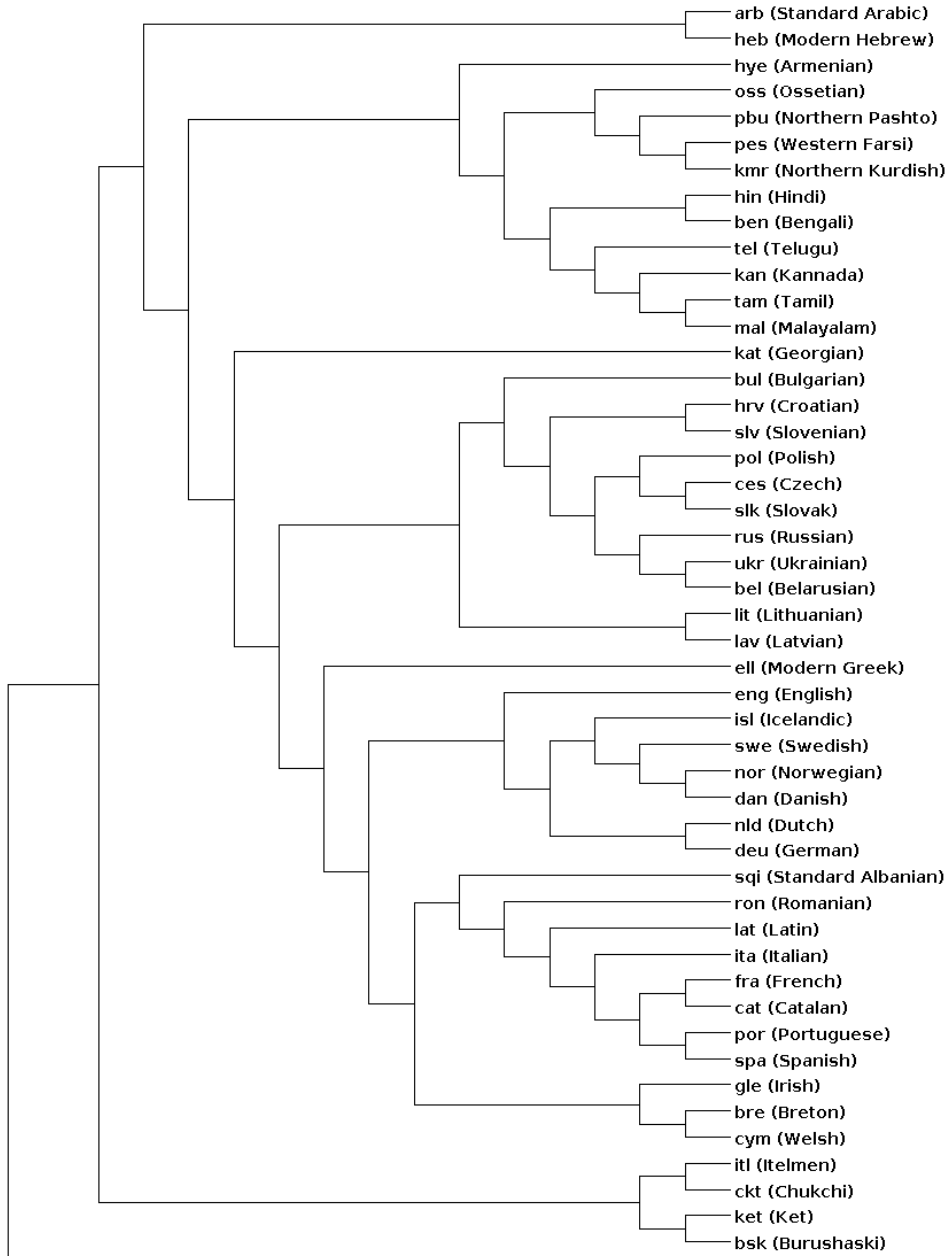


Fig. A.1.: Upper half of the language tree.

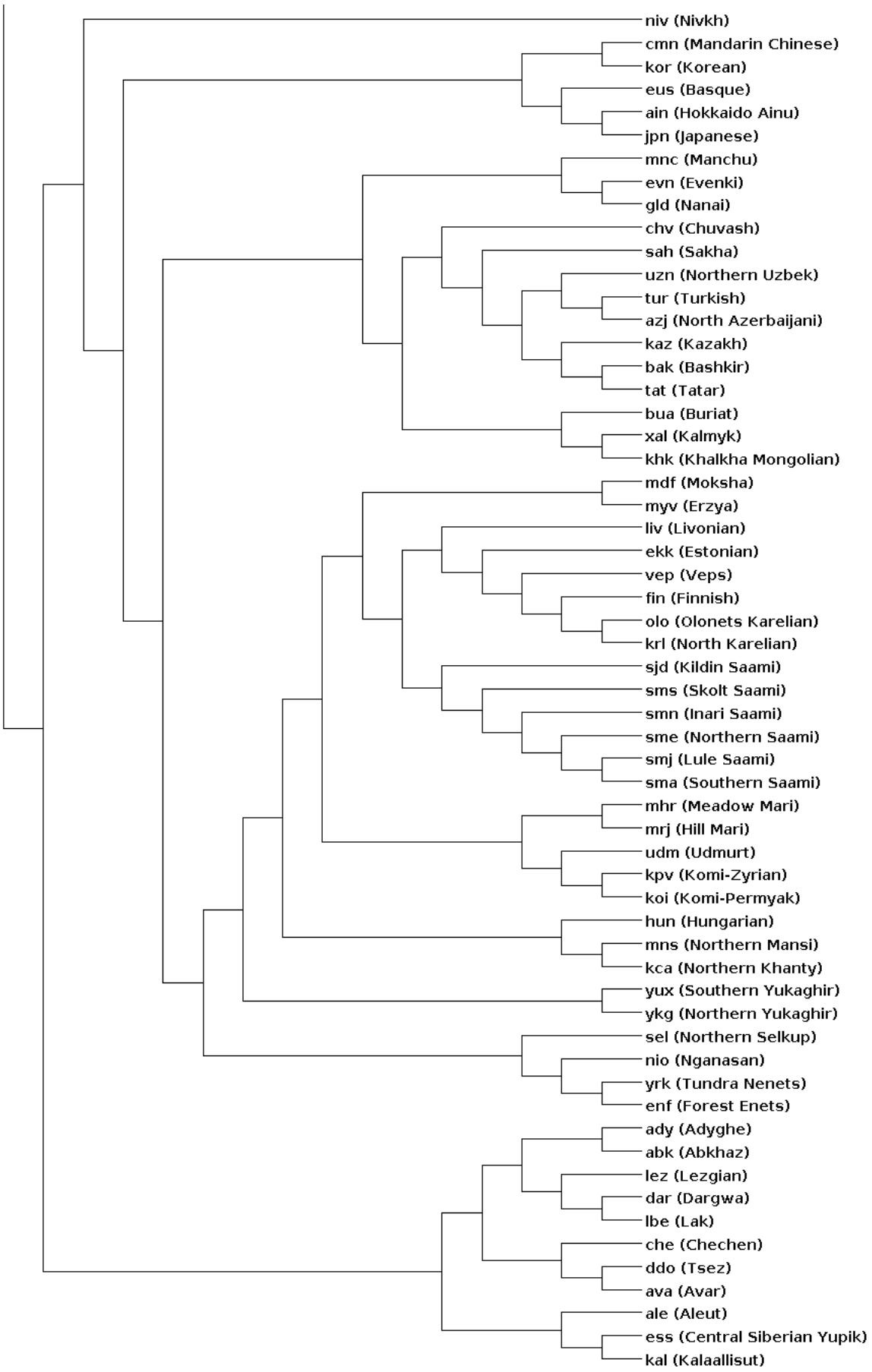


Fig. A.2.: Lower half of the language tree.

A.3 Sample Size Estimation for MLN Sampling

The MLN algorithm results in a language pair included in the transfer process, but cannot distinguish between RL and SL. In chapter 6.2.1, a workaround solution is presented to account for this issue. The procedure includes sampling of the data in order to estimate precision and recall for the MLNs. For each language pair present in a transfer event, the RL is chosen randomly. For the chosen RL, precision and recall are computed to maintain a statistical parameter. The sample size determines the number of sampling steps, and therefore the number of possible RLs used for the evaluation procedure. In order to obtain an overall estimate, the mean of all precision, recall, and F-score values is computed. As it can be seen in table A.2, the resulting numbers for precision, recall, and F-scores do not differ significantly. For a matter of running time, a sample size of $k = 100$ is chosen for the final analysis.

Sample size	Mean precision	Mean recall	Mean F-score
100	0.145	0.567	0.231
500	0.145	0.572	0.232
1,000	0.145	0.572	0.232

Tab. A.2.: Precision–recall values for the different sample sizes for MLN sampling.

A.4 Evaluation of Bootstrap Thresholds

A detailed representation of the evaluation scores for each bootstrap threshold is given in table A.3a for the tree comparison algorithm, and in table A.3b for the BD-based HGT algorithm. Since the F-scores are not illustrated in the precision–recall curve, they are listed in the corresponding tables. It can clearly be seen that independent of the thresholds, the recall decreases, whereas the F-score does not increase, which is an indicator of a great amount of noise present in the results of the tree reconciliation algorithms.

Method	Threshold	Precision	Recall	F-score
$CT_{Bayesian+NW}$	0.1	0.152	0.424	0.224
	0.2	0.151	0.414	0.221
	0.3	0.152	0.404	0.221
	0.4	0.153	0.391	0.220
	0.5	0.150	0.360	0.212
	0.6	0.156	0.333	0.212
	0.7	0.156	0.288	0.202
	0.8	0.155	0.230	0.186
	0.9	0.156	0.164	0.160
$CT_{ML+ngramsNW}$	0.1	0.148	0.395	0.215
	0.2	0.150	0.391	0.217
	0.3	0.153	0.370	0.217
	0.4	0.160	0.326	0.215
	0.5	0.160	0.244	0.193
	0.6	0.168	0.165	0.167
	0.7	0.211	0.102	0.138
	0.8	0.260	0.042	0.073
	0.9	0.371	0.010	0.0212
CT_{dow}	0.1	0.167	0.344	0.225
	0.2	0.172	0.336	0.228
	0.3	0.175	0.314	0.224
	0.4	0.176	0.288	0.219
	0.5	0.181	0.264	0.215
	0.6	0.186	0.225	0.204
	0.7	0.199	0.197	0.198
	0.8	0.222	0.155	0.183
	0.9	0.227	0.099	0.138

(a) Precision–recall values for the bootstrap thresholds for the tree comparison algorithms.

Method	Threshold	Precision	Recall	F-score
$CT_{Bayesian+NW}$	0.1	0.143	0.537	0.227
	0.2	0.142	0.404	0.210
	0.3	0.143	0.321	0.198
	0.4	0.145	0.252	0.184
	0.5	0.143	0.198	0.166
	0.6	0.135	0.143	0.139
	0.7	0.138	0.111	0.123
	0.8	0.130	0.073	0.093
	0.9	0.150	0.040	0.063
$CT_{ML+ngramsNW}$	0.1	0.150	0.243	0.185
	0.2	0.158	0.148	0.153
	0.3	0.160	0.087	0.113
	0.4	0.178	0.058	0.088
	0.5	0.166	0.034	0.057
	0.6	0.157	0.017	0.031
	0.7	0.226	0.014	0.026
	0.8	0.230	0.007	0.014
	0.9	0.294	0.004	0.008
CT_{dow}	0.1	0.143	0.653	0.235
	0.2	0.151	0.542	0.236
	0.3	0.163	0.462	0.241
	0.4	0.168	0.365	0.230
	0.5	0.173	0.270	0.211
	0.6	0.176	0.199	0.187
	0.7	0.176	0.147	0.160
	0.8	0.182	0.101	0.129
	0.9	0.192	0.061	0.093

(b) Precision–recall values of the bootstrap thresholds for the BD-based HGT algorithms.

Fig. A.3.: Evaluation of the bootstrap thresholds for the sequence-based algorithms.

A.5 Availability of the Programming Code

The programming code is available on my GitHub profile MarisaKoe (`https://github.com/marisaKoe`), where the repositories are organized in a project named *Automatic Loanword Identification using Tree Reconciliation* (`https://github.com/users/marisaKoe/projects/1`). In the following, the names of the programs are allocated to the chapters of the dissertation.

Chapter	Description	URL GitHub
Chapter 4.1.1	Reconstruction of state-of-the-art distance-based concept trees	`https://github.com/marisaKoe/conceptTreesPMI`
Chapter 4.1.2	Reconstruction of distance-based concept trees using cognate data	`https://github.com/marisaKoe/conceptTreesCognateData`
Chapter 4.1.3	Combining sequence and geographical distances for concept tree reconstruction	`https://github.com/marisaKoe/geo_conceptTrees`
Chapter 4.1.4	Reconstruction of the language tree using Bayesian inference	`https://github.com/marisaKoe/languageTreeMrBayes`
	Standard Bootstrapping for the language tree	`https://github.com/marisaKoe/languageTreePMI`
	Noisy bootstrapping for concept trees	`https://github.com/marisaKoe/bootstrappingWithNoise`
Chapter 4.2	Reconstruction of binary matrices and character-based tree reconstruction	`https://github.com/marisaKoe/conceptTreesCharBased`
Chapter 4.2.3	Reconstruction of the binary data matrices using substring sequences	`https://github.com/marisaKoe/binaryDataMatrices`
Chapter 4.1.5 and 4.2.4	Extract the MCC tree from the tree replicates	`https://github.com/marisaKoe/mccTree`
	Evaluation of the most reliable method according to the MCC trees	`https://github.com/marisaKoe/mccEvaluation`
Chapter 5.2.3	Tree rooting using MAD	`https://github.com/marisaKoe/rootingMAD`
Chapter 5.1.1	Horizontal Language Transfer using BD-based HGT algorithm	`https://github.com/marisaKoe/loanwordDetection-HGT`