Experimental explorations of a discrimination learning approach to language processing

D i s s e r t a t i o n
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Petrus Hendrikus Gertrudis Hendrix

aus

Grubbenvorst, den Niederlanden

2016

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Harald Baayen

Mitberichterstatter: Prof. Dr. Detmar Meurers

Mitberichterstatter: Prof. Dr. Gary Libben

Tag der mündlichen Prüfung: den 15. Dezember 2015

To my parents,

And to June,

Acknowledgements

This dissertation would not have been possible without the help and support of many people. I would like to take a moment to express my gratitude towards these people, who I hope will continue to be a part of my academic and personal life for many years to come.

First, I thank my thesis supervisor, Harald Baayen. Over the years, Harald's immense knowledge and endless academic creativity have been a continuous source of inspiration. I am extremely fortunate to have had the privilege of learning the psycholinguistic trade from one of the most innovative minds in our field today.

Apart from his academic guidance, I would also like to thank Harald for his personal support. His calm and positive attitude has been of great help, particularly in the later half of my doctoral student life. Furthermore, his kind-heartedness and generosity have allowed me to not only pursue my academic career, but to also enjoy the comfort of a stable family life. For this I cannot possibly thank you enough, Harald.

Harald also introduced me to Michael Ramscar. Michael's work on discrimination learning forms the conceptual foundation of this dissertation. His wisdom with regards to (the behavioral consequences of) discrimination learning principles has afforded me a much deeper understanding of the topic than would otherwise have been possible. In addition, I thank Michael for paving my way towards an internship at Facebook. My time at Facebook was a wonderful experience that provided me with valuable insights into how academic knowledge is applied in non-academic environments. Most of all, however, I thank Michael for being the amazing person he is.

Furthermore, I am greatly indebted to Detmar Meurers and Gary Libben for finding time in their busy schedules to review my dissertation. Detmar's comments provided me with valuable new insights into the challenges that lie ahead for a discrimination learning approach to language processing. Gary's higher-level thoughts helped me obtain a broader perspective on the work presented here. I also thank Andrea Weber and Petar Milin for being on the committee for my oral defense and their interesting questions during this defense. I specifically thank Petar Milin for his refreshing ideas about the properties of discrimination learning networks that drive language processing.

I am obliged to Patrick Bolger for his collaboration on the ERP study described in Chapter 4 of this dissertation and to Jonas Nick, who helped prepare the data frame for and was involved in the early stages of the analyses of the compound reading analysis presented in Chapter 3. Marco Marelli has been a joy to work with, both during his visits Edmonton and Tübingen, as well as when our research group visited Rovereto. Antal van den Bosch provided me with helpful advice for my stay at Facebook and kindly invited me to teach a statistics workshop in Nijmegen. I gained a better understanding of the eye-tracking data presented in Chapter 3 of this dissertation thanks to the insightful feedback of Reinhold Kliegl. I would also like to thank Antoine Tremblay for his great work on the ERP analysis technique used in Chapter 4 of this dissertation.

Next, I would like to thank the members of the Quantitative Linguistics research group. Tineke drastically reduced the amount of time I had to spend on administrative issues and always responded to my requests in an equally timely and cheerful manner. Cyrus kindly and adequately answered my countless technical questions and provided a reliable high performance computing environment, without which many of the simulations in this dissertation would have been impossible. Jacolien generously shared her thorough knowledge of GAM(M)s whenever I knocked on her door. Lea and Tino translated the summary of this dissertation to German. I would like to thank the Quantitative Linguistics group in its entirety for providing a research environment that was both extremely inspiring and a lot of fun. Arvi, Dennis, Fabian, Jianqiang,

Koji, and all the Hiwi's that have been a part of our group over the last years, I thank you all for a superb time.

Finally, I thank my family. I am much obliged to my parents. They have always supported me and have encouraged me to pursue my personal and academic dreams. Mom, dad, thanks for everything. I also thank my brothers, Bart and Danny, for the time we have been able to spend together despite our very different life schedules. Means a lot to me.

June, Delight and Hope: you have given me a place I call home. Delight and Hope, I am very thankful for welcoming me into your family. I could honestly not wish for more wonderful children than the two of you. June, thank you for the endless energy you invested in providing me with the optimal conditions to write this dissertation. I am blessed to have met a wonderful woman like you. I love you.

Contents

List of Tables

List of Figures

1 Introduction

"Data! Data! Data!" he cried impatiently.
"I can't make bricks without clay."

Sir Arthur Conan Doyle
The Adventures of Sherlock Holmes (1892)

Language is a fascinating phenomenon. Without much effort, we communicate with others about the world we live in. We know how to refer to countless objects and events using many different words and phrases, and understand others when they provide us with similar linguistic symbols. How does this impressive communication system come into place? How do we learn which linguistic symbols describe which objects and events? What insights does the way we acquire this knowledge provide us about how we use language as adults?[1]

This dissertation is an attempt to better understand the adult language processing system by investigating it from a learning perspective. In three chapters, I evaluate the performance of a language processing

[1] (see Ramscar et al., 2010, for a comprehensive discussion of these issues)

model that is based on a general human learning algorithm. The three chapters make use of three different psycholinguistic data sets: word naming latencies, eye-movements patterns during compound reading and the ERP signal in a picture naming task. Before turning to these data sets, however, I introduce the discrimination learning model that forms the conceptual and computational core of the language processing model proposed here. Furthermore, I provide a short overview of some of the applications of similar learning models in psycholinguistic studies that inspired the work presented here, as well as in recent studies that I contributed to, but that are not included integrally in this dissertation.

1.1 Learning language

Language is based on symbolic thought. Symbols like words, signs or pictures shape the way in which we perceive and communicate about the objects and events in the world around us. An important question for computational models of language processing, therefore, is how linguistic symbols are represented in our mind. Broadly speaking, there are two types of language processing models: symbolic models and sub-symbolic models.

As noted by Chalmers (1992), in symbolic theories of language learning, each computational unit represents a discrete linguistic symbol – such as a letter or a word. The level of computation and representation, therefore, is the same in symbolic models of language learning. By contrast, sub-symbolic theories hold that the level of computation is lower than the level of representation. Rather than being represented as discrete computational units, linguistic symbols are represented as activation patterns of sets of computational units. In other words: in symbolic models the words "table" and "chair" are represented by different computational units, whereas in sub-symbolic models these words are represented by different activation patterns of the same computational units (see Chalmers (1992) for a comprehensive discussion of the differences between symbolic and sub-symbolic approaches).

The most typical example of sub-symbolic models of language processing are connectionist models (cf. Fodor & Pylyshyn, 1988; Rumelhart & McClelland, 1986). Connectionist models have had a substantial influence in computational psycholinguistics over the last decades and have been applied with considerable success in a variety of psycholinguistic studies (see, e.g., Seidenberg & McClelland, 1989; Harm & Seidenberg, 2004). Nonetheless, sub-symbolic approaches to language processing are associated with a number of problems. A first problem concerns the fact that connectionist models learn through back-propagation of error. In back-propagation learning, the weights between input units and output units are adjusted based on a comparison of the model output to the target output. As noted by Perry et al. (2007), back-propagation learning is implausible from a neurobiological perspective (see, e.g., Crick, 1989; Murre et al., 1992; O'Reilly, 1998, 2001). Second, sub-symbolic models tend to be less transparent than symbolic models. The interpretability of activation patterns over a large set of units tends to be reduced as compared to the interpretability of activations of individual units. Furthermore, most connectionist models are multi-layer networks, in which a layer of hidden units connects the input units and the output units (see, however, Harm & Seidenberg, 2004 for a two-layer connectionist network without hidden layer units). Layers of hidden units reduce the transparency of sub-symbolic models of language processing.

Symbolic models of language learning are associated with problems of their own. Most prominently, discrete linguistic symbols are an oversimplification of reality. As pointed out by Ramscar et al. (2010, p. 911), "the kinds of things that people represent and think about symbolically do not fall into discrete classes, or categories, of Xs and Ys [...]; symbolic categories do not possess discrete boundaries (i.e., there are no fixed criteria for establishing whether an entity is an X or a Y); and entities are often assigned to multiple symbolic classes (i.e., they are sometimes Xs; sometimes Ys)". As a result, representations in symbolic models of language learning are by definition limited to approximations of more complex representations in our mental lexicon.

Nonetheless, symbolic models of language learning tend to provide excellent learning performance and increased transparency as compared to sub-symbolic language processing models. The increased transparency of symbolic models offers important insights into the probabilistic patterns in the linguistic environment that drive behavioral patterns in psycholinguistic experiments. In my attempt to better understand the adult language processing system by investigating it from a learning perspective, I therefore opted to use a simple two-layer symbolic network model of language learning.

1.2 Discrimination learning

The learning model adopted here is a discrimination learning model: the Rescorla-Wagner model (Rescorla & Wagner, 1972). The description of discrimination learning below is an adapted version of the presentation of the naive discrimination learning approach in Baayen et al. (2011) and Baayen et al. (2013). For further information we refer the interested reader to these papers.

The Rescorla-Wagner equations describe a two-layer symbolic network model that operates on the basis of cues, outcomes and the associations between them. Formally, the association strength V_i^{t+1} between a cue C_i and an outcome O at time $t+1$ is defined as:

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{1.1}$$

with:

$$\Delta V_i^t = \begin{cases} 0 & \text{if} \quad \textsc{absent}(C_i, t) \\ \alpha_i \beta_1 \left(\lambda - \sum_{\textsc{present}(C_j, t)} V_j \right) & \text{if} \quad \textsc{present}(C_j, t) \ \& \\ & \quad \textsc{present}(O, t) \\ \alpha_i \beta_2 \left(0 - \sum_{\textsc{present}(C_j, t)} V_j \right) & \text{if} \quad \textsc{present}(C_j, t) \ \& \\ & \quad \textsc{absent}(O, t) \end{cases} \tag{1.2}$$

The parameters α and β refer to the salience of the cue and outcome, respectively. By default, all α's are equal, and $\beta_1 = \beta_2$. The parameter λ refers to the maximum association strength and is typically set to 1.

As can be seen in Equation 1.2, the association strength between a given cue and outcome is modified in two cases. First, whenever a cue and outcome occur together, the association strength between the cue and outcome increases. Second, when a cue occurs in the absence of an outcome, the association strength between a cue and outcome decreases. As such, "the learning process is driven by discrepancies between what is expected and what is observed" (Ramscar et al., 2010, p. 913). The Rescorla-Wagner equations are therefore described as an instantiation of error-driven learning.

The idea that learning is driven by the differences between predictions and expectations is well-established in the learning literature. Gallistel (2003), for instance, argues that learning is possible only when the observed entropy of an event diverges from the maximum entropy. The concept of error-based learning is further supported by a number of neurobiological studies. Waelti et al. (2001), for instance, showed that dopamine responses for expected and unexpected outcomes in monkeys comply with the predictions of error-driven learning models (see, e.g., Daw & Shohamy, 2008; Schultz, 2002; Hollerman & Schultz, 1998) for further neurobiological studies on error-driven learning).

For any given linguistic input only a small set of cues is present in the input. The activation a_k of an outcome O_k given the set of cues in the input C is defined as:

$$a_k = \sum_{j \in C} V_{jk}. \tag{1.3}$$

with j ranging over the active cues and V_{jk} being the association strength between cue C_j and outcome O_k. This activation is a measure of the amount of bottom-up support for outcome O_k given cues C and is typically inversely proportional to processing times, such that more bottom-up support corresponds to faster processing latencies.

As noted by Baayen et al. (2011), the instantiation of error-based learning in the Rescorla-Wagner model is one that has a rich tradition in the cognitive psychology literature (cf. R. R. Miller et al., 1995; Siegel & Allan, 1996). The model is highly similar to a perceptron (Rosenblatt, 1958) and to the delta rule (Widrow & Hoff, 1960). In a broader sense, Baayen et al. (2011) continue, the Rescorla-Wagner equations can be thought of as a general-purpose probabilistic learning algorithm (see Chater et al., 2006; Hsu et al., 2010). Baayen (2011a), for, instance, demonstrated that the Rescorla-Wagner equations perform as well as state-of-the-art machine learning techniques for predicting the dative alternation in English.

Recently, the Rescorla-Wagner learning model has been applied in a number of psycholinguistic studies. For child language acquisition, Ramscar and Yarlett (2007) found that the typical U-shaped development that characterizes the acquisition of irregular plural forms is predicted by a discrimination learning model (see Ramscar, Dye & McCauley, 2013). Hsu et al. (2010) showed that in first language learning many of the phenomena typically attributed to innate language-specific biases are alternatively explained by prediction-based learning. For second language learning, Ellis (2006) demonstrated that many of the problems and limitations that second language learners encounter are predicted by discrimination learning.

1.3 The adult language processing system

The language acquisition studies mentioned above simulate the time-course of learning through the iterative Rescorla-Wagner equations. Initially, the association strengths in the Rescorla-Wagner model are subject to substantial fluctuations as a function of linguistic experience. The results from the studies above indicate that these initial fluctuations provide a good characterization of the learning process in language acquisition. Over time, however, the association strengths in the Rescorla-Wagner model asymptote towards an equilibrium state. This equilibrium state can be used as a proxy for the adult language processing system.

The equilibrium state of the Rescorla-Wagner model is formalized in the equilibrium equations provided by Danks (2003). The equilibrium equations, as implemented in the NDL R package (Shaoul, Arppe et al., 2013), define the connection strength (V_{ik}) between cue (C_i) and outcome (O_k) as:

$$\Pr(O_k|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i)V_{jk} = 0 \tag{1.4}$$

where $\Pr(C_j|C_i)$ is the conditional probability of cue C_j given cue C_i, $\Pr(O_k|C_i)$ is the conditional probability of outcome O_k given cue C_i and $n+1$ is the number of different cues.

For computational efficiency, the association strengths from cues to a specific outcome O_k are estimated separately and independently of all other outcomes. This assumption of independence is a simplification of reality that is conceptually similar to the independence assumption in the naive Bayes classification algorithm and inspired Baayen et al. (2011) to refer to the discrimination learning algorithm used throughout this dissertation as naive discrimination learning.

In Baayen et al. (2011), we first used the equilibrium equations for the Rescorla-Wagner model to explore the adult language processing system. The naive discrimination learning (henceforth NDL) model in Baayen et al. (2011) takes as input cues letters and letter combinations and as outcomes lexico-semantic representations. The model is a full-decomposition model of morphological processing, in which there are no separate representations for morphologically complex words. We demonstrated that this discrimination learning model captures a wide range of effects observed in the experimental psycholinguistic literature, including morphological family size effects, inflectional entropy effects, constituent and whole-form frequency effects for complex words and paradigmatic entropy effects.

In a follow-up study (Baayen et al., 2013), we found that the same full-decomposition NDL model captures the phrase frequency effects described by Arnon and Snider (2010). Arnon and Snider (2010) demonstrated that

phrasal decision latencies for frequent phrases such as "all over the place" were shorter as compared to phrasal decision latencies for infrequent phrases such as "all over the city". These effects could not be reduced to frequency effects of single words or component n-grams.

As noted by Baayen et al. (2013), phrase frequency effects are often interpreted as evidence for phrase-level representations. This interpretation fits well with theories of language processing that assume storage of and computation over large numbers of stored phrase-level representations, such as data-oriented parsing (Bod, 2006) and memory-based learning (Daelemans & Bosch, 2005). The NDL model, however, successfully captures the phrase frequency effect in Arnon and Snider (2010), without assuming any representations beyond the word level. As such, the NDL model provides a simpler and more economical account of phrase frequency effects as compared to storage-based alternatives.

1.4 Cognitive aging

A particularly fruitful application of naive discrimination learning in the context of the adult language processing system concerns the topic of cognitive aging. In a series of papers, we investigated the consequences of additional experience with a language on the performance in a variety of linguistic tasks. By providing an NDL model with varying amounts of linguistic input, we demonstrated that many of the findings typically attributed to cognitive decline are a straightforward consequence of increased linguistic experience.

A typical finding in the experimental psycholinguistic literature, for instance, is that older people have longer lexical decision latencies than younger people, particularly for low frequency words. This age by frequency interaction follows straightforwardly from a discrimination learning model. High frequency words are encountered regularly, which leads to a constant reinforcement of the associations between the letters and letters combinations in these high frequency words and their lexemes. Low frequency words, however, are encountered on a much less regular basis.

As noted by Baayen (2014), low frequency words are initially "protected" by the fact that they tend to consist of less frequent letter combinations than high frequency words. The word "qatar", for instance, contains the low frequency word-initial letter bigram "qa". For many young participants, "qatar" may be the only word with a word initial "qa" bigram in their mental lexicon. For these participants, "qa" is an excellent cue for the lexeme "qatar", which allows for relatively fast lexical decision latencies. Older participants, by contrast, may have experienced other words that start with "qa", such as "qanat" (i.e., "a gently sloping underground channel or tunnel constructed to lead water from the interior of a hill to a village below"), "qat" (i.e., "the leaves of an Arabian shrub, which are chewed as a stimulant" or "qaid" (i.e., "Muslim tribal chief"). For these participants, the letter combination "qa" is a less reliable cue for the word "qatar", which results in longer lexical decision latencies (example adapted from Baayen, 2014).

In simulations reported in Ramscar et al. (2014), we demonstrated that an NDL model replicates the age by frequency interaction described above for the lexical decision data made available by Balota et al. (1999) when provided with different amounts of linguistic input that reflect the linguistic experience of older and younger participants. In other words: using the exact same computational architecture and hardware, increased experience with a language leads to exactly the type of problems with low frequency words that we observe in older participants.

A second finding in the experimental psycholinguistic literature that is typically attributed to cognitive decline, Ramscar et al. (2014) continue, is the decreased performance of older participants in the paired associate learning task. In paired associate learning participants are asked to memorize pairs of words such as "north-south" or "jury-eagle". In the recall phase the first word of a pair is presented (e.g., "north", "jury") and participants are asked to produce the second word (e.g., "south", "eagle"). Older people perform less well on this task as compared to younger people. This age effect is more prominent for harder word pairs such as "jury"-"eagle" than it is for easier word pairs such as "north"-"south".

Again, Ramscar et al. (2014) argue, cognitive decline is an unlikely explanation for the decreased performance of older people in the paired associate learning task. Logically, older participants might perform worse than younger participants due to a general decrease in cognitive ability as a function of age. If this interpretation of the age effect were correct, however, it is unclear why older people should perform worse on hard word pairs, but not on easier word pairs. From a discrimination learning perspective, by contrast, we expect exactly the type of age by item difficulty interaction that is typically observed in paired associate learning studies.

While the experience of older participants with the words "north" and "south" may be somewhat more diverse, both younger and older people are aware of the high co-occurrence rate of these words. For both groups of participants "north" is an excellent cue for "south". Older and younger participants therefore show comparable performance for items like "north-south". In discrimination learning models, however, expectations are shaped not only by positive, but also by negative evidence. Through a lifetime of linguistic experience older participants have learned that words like "north" and "south" often occur together, but also that words like "jury" and "eagle" do not. The "decreased performance" of older participants for harder items in paired associate learning, Ramscar et al. (2014) argue, might therefore reflect increased awareness of the fact that the word "jury" is an uninformative cue for the word "eagle".

In Sun et al. (2015), we demonstrate that a re-analysis of the Rosiers and Ivison (1986) normative data using a (beta regression) generalized additive mixed-effect model supports such an interpretation of the age by item difficulty interaction in paired associate learning. Figure 1.1[2] shows this interaction, with warmer colors representing poorer performance. As can be seen in Figure 1.1, the performance of older participants for pairs with a high co-occurrence frequency is on par with that of younger participants. Only when the words in a pair do not occur together very often, we see a clear "decrease" in performance for older participants. As such, older participants show increased sensitivity to the co-occurrence

[2] This figure is a slightly adapted version of the figure in Sun et al. (2015)

frequency of the words in a pair. This pattern of results is hard to explain from a cognitive decline perspective (which would expect a decrease in performance for older participants across the co-occurrence frequency range), but follows straightforwardly from the principles of discrimination learning outlined above.

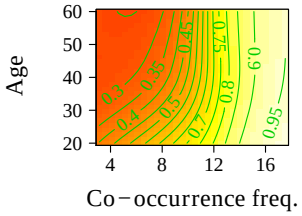

Figure 1.1. Effects of (log-transformed) Co-Occurrence Frequency by Age on Paired Associate Learning performance.

A third phenomenon interpreted as evidence for cognitive decline is the fact that older people have problems remembering names and retrieving names from memory (see, e.g., G. Cohen & Faulkner, 1986). Through a series of simulations (Ramscar et al., 2014; Ramscar, Smith et al., 2013), we show that these name retrieval problems are a natural consequence of two aspects of changes in the linguistic distribution of names over the lifetime. First, the perplexity of the English name system has increased almost exponentially over the last 50 years. The processing load imposed by names has therefore increased dramatically during the lives of older people. Second, older people have experienced more names simply because they are older and have met and communicated about greater numbers of people. Even if we do not assume cognitive decline, therefore, older people should be expected to have a harder time remembering and retrieving names.

The results described above demonstrate that many of the experimental findings that are typically interpreted in terms of cognitive decline can alternatively be explained by increased experience with the language. Indeed, in simulations reported in Ramscar, Hendrix et al. (2013) and Ramscar et al. (2014), we demonstrate that it is exactly this increased

experience that leads to improved performance of older people in experimental paradigms like the FAS subtest of the Controlled Oral Word Association Test (Spreen & Strauss, 1998), where people are asked to generate words starting with the letters "F", "A" or "S". Depending on the nature of the task, therefore, increased linguistic experience can lead to either improved or decreased performance in tests of cognitive ability.

The studies on cognitive aging discussed above do not attempt to provide conclusive evidence against the notion of cognitive decline. It is well possible that certain aspects of human behavior inevitably suffer from a general decrease in cognitive functioning over the lifetime. For the lexical learning tasks described above, however, there is surprisingly little evidence for age-related cognitive decline once linguistic experience is controlled for. Instead, the behavior of older participants in a variety of tasks suggests an increased sensitivity to the distributional properties of the language. As such, traditional views on cognitive aging may considerably overestimate the degree to which cognitive functioning declines with age. A more informed understanding of (lexical) processing over the lifetime that takes into account the consequences of learning will help better understand the costs and benefits associated with cognitive aging.

1.5 Data, data, data!

This dissertation began with a citation from The Adventures of Sherlock Holmes (Doyle, 1982): "'Data! Data! Data!', he cried impatiently. 'I can't make bricks without clay.'". Ever since Harald Baayen introduced me to the work of Michael Ramscar, I have thought of discrimination learning as an exciting and promising new approach to language processing. Having the privilege of being involved in the first explorations of the adult language processing system from a (naive) discrimination learning perspective in a collaboration with Harald Baayen, Petar Milin, Dušica Filipović Durdević, and Marco Marelli (Baayen et al., 2011) further increased my enthusiasm for the subject. The successful simulation of many of the effects observed in the experimental morphological literature in Baayen et al. (2011) resulted in an almost irresistible urge to further

explore the explanatory power of a discrimination learning approach for a variety of psycholinguistic data sets. "Data! Data! Data!".

In the three chapters that follow this introductory chapter, I will describe the results of an evaluation of the naive discrimination learning (henceforth NDL) approach to adult language processing for three different psycholinguistic data sets. Each data set evaluates the performance of the NDL approach for a different psycholinguistic measure and a different linguistic phenomenon. Chapter 2 presents an NDL model that simulates reaction times for the reading aloud of monosyllabic words and non-words. Chapter 3 gauges the explanatory power of NDL measures for eye-fixation patterns on noun-noun compounds in a new large-scale database of eye-movements that were recorded while participants read a collection of fictional texts, the Edmonton-Tübingen eye-tracking corpus (ET corpus). Chapter 4 evaluates the performance of an NDL approach in accounting for the ERP signal in a primed picture naming task.

Taken together, Chapters 2, 3 and 4 highlight the potential of a naive discrimination learning approach for understanding the adult language processing system for a variety of behavioral measures and experimental paradigms, and bring to light some of the limitations of the naive discrimination learning approach and the implementation of this approach in the NDL simulations described here. In the final chapter of this dissertation I will briefly evaluate the overall performance of the NDL approach for the data sets presented here. Furthermore, I will outline some of the challenges that lie ahead in future research that adopts a discrimination learning approach to the adult language processing system.

2

Word naming

2.1 Introduction

Both M. Coltheart et al. (2001) and Perry et al. (2007) open what have become canonical papers in the reading literature with the observation that tremendous advances have been made in the development of reading models over the last decades. They note note that early cognitive models in psychology provided mainly verbal descriptions of hypothesized cognitive architectures. These models took the form of flowchart diagrams in which boxes were used to depict mental representations, which were manipulated by cognitive processes represented as arrows that connected the various boxes (see, e.g., Morton, 1969) for an application of box-and-arrow models to reading). Although such "verbal" models provide *descriptions* of behavioral data, their lack of specificity meant that they could only be related to the psychological and neurobiological reality of language processing at a very abstract level.

The recent development of more formal, computationally implementable models of reading (see, e.g., M. Coltheart et al., 2001; Harm & Seidenberg, 2004; Perry et al., 2007, 2010) has done much to address this shortcoming. As M. Coltheart et al. (2001) remark, the development of a computational model requires a precise specification of any processes

and representations that are to be implemented. As a result, computational models offer a clear improvement in specificity over informal "verbal" models of reading. Because computational models generate precise and explicit predictions, M. Coltheart et al. (2001) continue, it is possible to evaluate them against existing behavioral data, and even falsify them through later findings. In addition, recent advances in cognitive and computational neuroscience have provided opportunities to complement this approach with even more stringent tests that investigate the neurobiological plausibility of a model's architecture and processing mechanisms.

Since the initial implementation of the Dual-Route-Cascaded model by M. Coltheart et al. (2001), a decade-and-a-half of recursive implementation and assessment of computational models have provided valuable insights into the successes of and the challenges for models of reading aloud. Consequently, the qualitative and quantitative performance of current state-of-the-art models of reading aloud is orders of magnitude better than that of previous generations of models.

Although current state-of-the-art models of reading aloud (see, e.g., M. Coltheart et al., 2001; Harm & Seidenberg, 2004; Perry et al., 2007, 2010) differ with respect to the exact mechanisms they propose, they all divide the process of reading aloud into two "routes" (i.e., sub-processes). The first route is a "lexical route", in which mappings from orthography to phonology are mediated by lexical representations. This allows the reading of known words such as "wood" and "blood" to be simulated. The second route is a "sub-lexical" route that directly maps orthographic units onto phonological units and allows for the simulation of reading potentially unknown words, such as "snood". As such, the general consensus seems to be that reading aloud is best modeled through a dual-route architecture. To cite M. Coltheart et al. (2001, p.303), "Nothing ever guarantees, of course, that any theory in any branch of science is correct. But if there is no other theory in the field that has been demonstrated through computational modeling to be both complete and sufficient, resting on laurels is a reasonable thing to do until the emergence of such

a competitor - that is, the emergence of a different theory that has also been shown to be both complete and sufficient."

In what follows, we hope to breath new life into the single versus dual-route debate by presenting a new single-route model of reading aloud, the Naive Discriminative Reading aloud (NDR$_a$) model. The NDR$_a$ is an extension of the NDR model for silent reading by Baayen et al. (2011), in which both words and non-words are read through a single lexical architecture. Following the fruitful tradition described above, we will evaluate the performance of the NDR$_a$ model for a wide range of effects documented in the experimental word and non-word naming literature. We show that the NDR$_a$ successfully captures the linear and non-linear characteristics of these effects, as well as a hitherto unobserved frequency effect for non-words. We further demonstrate that the addition of a sub-lexical route to the NDR$_a$ is redundant, in that it does not improve the performance of the model.

2.2 Existing models

In the reading aloud task, participants are presented with printed words on a computer screen and asked to pronounce these words as quickly and accurately as possible. Orthography and phonology play an important role in this process. These roles are undisputed in all current models of reading aloud, which contain both orthographic and phonological representations in one form or another. The role of semantics has been subject to a little more debate. While previous single-route models of reading aloud mapped orthography directly onto phonology, however, the consensus in more recent models is that the orthography-to-phonology mapping is mediated by semantic representations at least some of the time. Dual-route models of reading aloud have posited that while non-words are read through a direct orthography-to-phonology mapping, reading real words involves lexico-semantic representations.

Below, we discuss some of the existing state-of-the-art models of reading aloud. First, the triangle model (see, e.g., Seidenberg & McClelland, 1989; Plaut et al., 1996; Harm & Seidenberg, 2004) will be introduced.

Next, we discuss the Dual-Route Cascaded Model (M. Coltheart et al., 2001). We conclude with a description of the model of reading aloud that currently yields the best simulation results: the Connectionist Dual Process model (Zorzi et al., 1998b; Perry et al., 2007, 2010).

2.2.1 The triangle model

The triangle model (see, e.g., Seidenberg & McClelland, 1989; Plaut et al., 1996; Harm & Seidenberg, 2004) is a model comprising of three levels of description: orthography, phonology and meaning. Mappings between these levels of description are implemented as three-layer connectionist networks. The architecture of the model is presented in Figure 2.1.

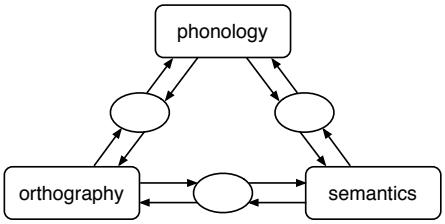

Figure 2.1. The basic architecture of the triangle model.

In the original version of the triangle model only the direct mapping from orthography to phonology was implemented (Seidenberg & McClelland, 1989). This original model therefore was a single-route model of reading aloud that directly mapped orthography onto phonology. Representations consisted of triplets of orthographic and phonological features (Wickelgren, 1969).[1] Associations between these orthographic and phonological units were learned through a 3-layer connectionist network.

Harm and Seidenberg (2004) added semantics to the triangle model. The latest version of the model therefore has two routes from orthography to phonology. The first route is a direct mapping from orthography to phonology, as in Seidenberg and McClelland (1989). In the second route the mapping from orthography to phonology is mediated by semantic

[1] These Wickelfeatures were replaced by more localist representations in Plaut et al. (1996)

representations. The addition of a second route to the model allowed Harm and Seidenberg (2004) to simulate a number of effects in the experimental literature that were not captured by previous versions of the triangle model, including effects of homophones and pseudo-homophones.

Being a connectionist model, the triangle model operates on the basis of a general learning mechanism. As such, the triangle model has increased plausibility over models that posit task-specific processing mechanisms (see Seidenberg, 2006). Connectionism, however, has its own share of disadvantages. First, most connectionist networks are multi-layer networks, in which the mapping between input and output units is mediated by one or more layers of hidden units.[2] The contents of these hidden layer units are opaque. This reduces the transparency and interpretability of connectionist models (see, e.g., Baayen et al., 2011). In addition, connectionist models learn through back-propagation of error. In back-propagation learning the model output is compared to the target output. The model weights are then updated on the basis of the difference between the model output and the target output (Rumelhart et al., 1986; Seidenberg, 2006). As noted by Perry et al. (2007), back-propagation learning has been criticized for being neurobiologically implausible (Crick, 1989; Murre et al., 1992; O'Reilly, 1998, 2001).

2.2.2 The Dual-Route Cascaded model

A different class of models was developed in parallel to the different versions of the triangle model. While later versions of the triangle model did include a second, lexical route (Harm & Seidenberg, 2004), the Dual-Route Cascaded model (henceforth DRC, M. Coltheart et al., 2001) was the first computational implementation of a dual-route architecture. The architecture of the DRC model is displayed in Figure 2.2.

The first stage of the model is shared by both routes and consists of an interpretation of the visual input in terms of visual features (Rumelhart & Siple, 1974) that activate letter units. From this orthographic level the phonological representations required for speech can be accessed

[2] Harm and Seidenberg (2004) implemented a 2-layer orthography to phonology mapping that does not contain hidden units.

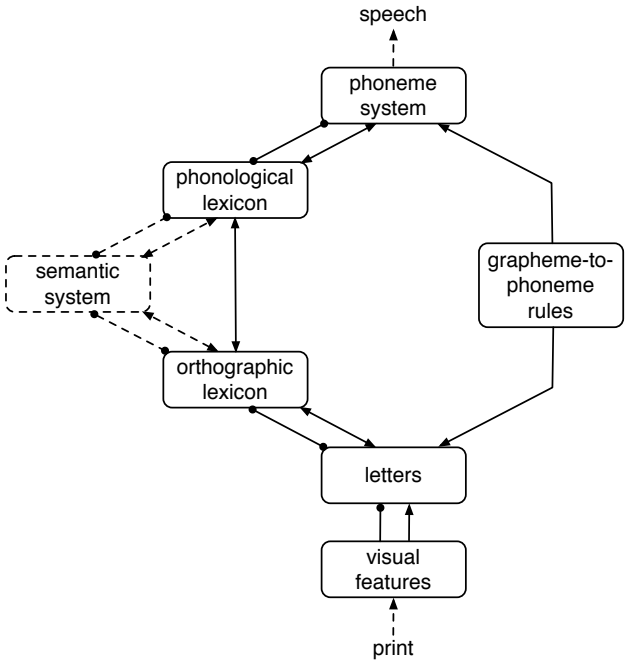

Figure 2.2. The basic architecture of the DRC model.

through two routes. The sub-lexical route maps letter units directly onto phonemes, whereas in the lexical route this mapping is mediated by a lexico-semantic system.

The sub-lexical route of the DRC model is based on grapheme-to-phoneme conversion rules (Rastle & Coltheart, 1999). This route, which is posited to be necessary for reading non-words, operates serially in an all-or-none fashion. This sub-lexical route also underlies the successful simulation of the increased processing costs associated with words with irregular orthography to phonology mappings (i.e., mappings not predicted by the set of rules in the model). As a result of the all-or-none operation of the grapheme-to-phoneme conversion rules, however, the model has problems simulating the results of graded consistency experiments in which the number and frequency of words with consistent (i.e., the same) and inconsistent (i.e., different) orthography-to-phonology mappings is

taken into account. Furthermore, the rule-based implementation of the sub-lexical route is psychologically and biologically less plausible than the learning algorithms that underlie the direct orthography to phonology mapping in other models.

The lexical route of the DRC model is based on the interactive activation model of McClelland and Rumelhart (1981) and is parallel rather than serial in nature. Like the rule-system in the sub-lexical route, the interactive activation model in the lexical route of the DRC model is fully hard-coded and ignores the problem of learning. A further problem is that it does not capture a number of important findings in the experimental literature (see, e.g., Andrews, 1996; Ziegler & Perry, 1998, see Perry et al., 2007 for a comprehensive discussion of the shortcomings of the DRC model).

2.2.3 The Connectionist Dual Process model

The latest dual-route model is the Connectionist Dual Process model (henceforth CDP, Zorzi et al., 1998b; Perry et al., 2007, 2010). Similar to the DRC, the different versions of the CDP model consist of a lexical and a sub-lexical route. The basic architecture of the CDP model is presented in Figure 2.3.

The major advancement of the CDP over the DRC model is the implementation of a two-layer associative learning network in the sub-lexical route (Zorzi et al., 1998b, 1998a). To learn the connection strengths between orthographic and phonological units the network uses the delta rule (Widrow & Hoff, 1960), which is a general algorithm for human learning (Siegel & Allan, 1996). As such, the implementation of the sub-lexical route of the CDP models is an important step towards a neurobiologically plausible model of reading aloud. In the most current versions of the CDP model this learning network was complemented with a graphemic buffer in the sub-lexical route (Perry et al., 2007). This graphemic buffer organizes the orthographic information into a graphosyllabic template that uses the most frequent graphemes as representational units (Perry et al., 2010; Houghton & Zorzi, 2003).

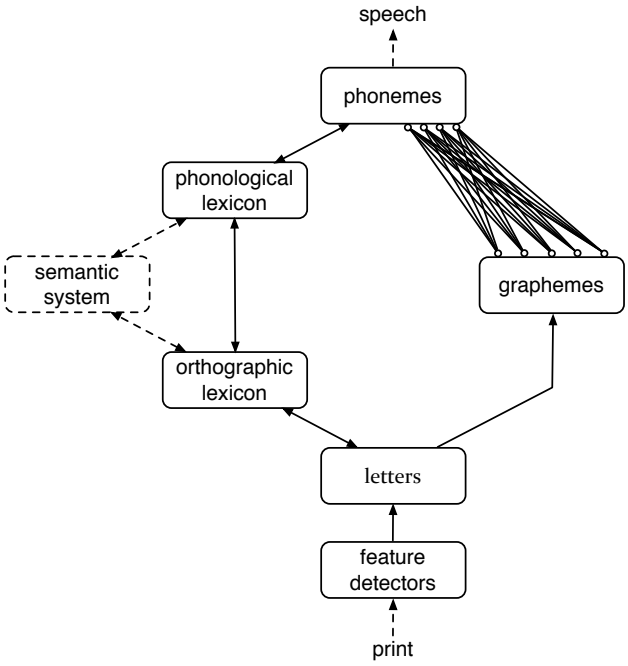

Figure 2.3. The basic architecture of the CDP+ model.

In the original CDP model the lexical route was, in the words of Perry et al. (2007, p.297), "not implemented beyond the provision of frequency-weighted lexical phonological activation" (see Zorzi et al., 1998b). The CDP+ model (Perry et al., 2007) implemented the lexical route of the DRC model to overcome this problem. In doing so, however, the latest versions of the CDP model inherited the problems of interactive activation models. As such, one of the problems of the CDP+ model is that there is no learning in the lexical route (see Perry et al., 2007, p.303).

The lexical and sub-lexical routes of the CDP+ model are connected at the orthographic input and phonological output levels. On the input side of the model the visual input (i.e., the printed word) is first decoded into features with a slightly altered version of the McClelland and Rumelhart (1981) feature detectors. These features are then translated into letters. At the output side of the model the information from the lexical

and sub-lexical routes is integrated in a phonological decision system. Naming latencies in the CDP+ model are based on a settling criterion that terminates processing when the network is in a stable state (see Zorzi et al., 1998b).

In a comprehensive study Perry et al. (2007) demonstrated that the CDP+ model accounts for a wide range of experimental findings and shows item-level correlations with observed naming latencies that are an order of magnitude higher than those in the DRC and the triangle model. We therefore consider the CDP+ model the leading model of reading aloud.

In a recent extension of the CDP+, Perry et al. (2010) extended the model to bi-syllabic reading aloud. This CDP++ model correctly captures a number of experimental effects that are specifically relevant for multi-syllabic words, including effects of stress and the number of syllables. For mono-syllabic words, the CDP++ model behaves very similar to the CDP+ model, with minor changes in parameter settings and the assignment of graphemes to slots in the graphemic buffer.

2.3 The Naive Discriminative Reading Aloud model

The Naive Discriminative Reading Aloud (NDR_a) model differs from existing models of reading aloud in two ways. First, the computational implementation of the NDR_a is entirely based on the general principles of human learning described by the Rescorla-Wagner equations (Rescorla & Wagner, 1972). These equations are similar to the delta rule that is used in the sub-lexical network of the CDP+ model. As such, the NDR_a stands in sharp contrast to the lexical route of the CDP+ model, which is based on the interactive activation model of McClelland and Rumelhart (1981). The computational engine of the NDR_a also differs substantially from the connectionist networks that underlie the triangle model. It uses simple, transparent two-layer learning networks that directly map input units onto output units. In contrast to connectionist networks, these networks do not rely on the often uninterpretable hidden layer units or

back-propagation of error. A detailed description of the Rescorla-Wagner learning principles was provided in Chapter 1 of this dissertation.

Second, unlike all of the models discussed in the previous section, the NDR$_a$ consists of a single lexical architecture. The most recent version of the triangle model and the DRC and CDP models assume the use of both a lexical and a sub-lexical route in reading aloud, whereas the earlier single-route implementations of the triangle model were sub-lexical in nature. By contrast, NDR$_a$ applies a single lexical mechanism in both word and non-word reading.

The architecture for word reading in the NDR$_a$ is straightforward and similar to the processes underlying word reading in the lexical routes of existing models. Visual stimuli activate orthographic units. These orthographic units activate lexical representations of target words. In addition, they spread activation to lexical representations of orthographically similar words. The lexical representations of both the target word and the orthographic neighbors then activate phonological output units.

We propose that the reading of non-words occurs in a similar fashion. For non-words, however, no lexical representations exist. Therefore, instead of activating the lexical representations of both the target word and orthographically similar words, non-word orthographies only activate the lexical representations of orthographic neighbors and only these lexical representations subsequently activate phonological units.

In what follows we demonstrate that a wide range of non-word reading effects documented in the experimental literature follow straightforwardly from this simple architecture. This architecture also accounts for a novel finding, namely a non-word frequency effect in reading aloud. This non-word frequency effect suggests that the distinction between words and non-words may not be as black and white as previously thought and provides independent evidence for the involvement of lexical processes in non-word reading, as well as for the single route architecture of the NDR$_a$.

2.3.1 Model architecture

The architecture of the NDR$_a$ model is presented in Figure 2.4. The model assumes that reading aloud involves three processing stages. In the first stage, the visual input is interpreted and decoded into orthographic units. In the second stage, these orthographic units activate lexical representations in the mental lexicon that we will refer to as lexemes (i.e., lexical targets that link orthographic, phonological and semantic properties of words (Aronoff, 1994). In the third stage these lexemes activate phonological output units. The second and third stages of the model are implemented as two-layer associative learning networks, using the Rescorla-Wagner learning rule.

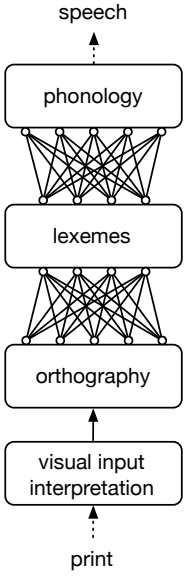

Figure 2.4. The basic architecture of the NDRa model.

2.3.2 Visual input interpretation

Prior to linguistic processing a decoding of the visual input is necessary. Both the DRC and the CDP+ use feature detection mechanisms that are similar in nature to the features detection mechanism in McClelland and Rumelhart (1981). The visual input interpretation mechanism in the NDR_a is a quantitative implementation of a feature decoding mechanism that is based on the idea that more complex visual patterns should take longer to decode.

We used a variant of the Manhattan city-block distance measure (see, e.g., Han & Kamber, 2000) to quantify the complexity of a letter in English. First, we constructed vector representations of the bitmaps of all 26 letters as written in black Lucida typewriter font on a white background (font size: 16). Each vector contained 400 elements representing the bit values for 20 horizontal and 20 vertical pixels. Black pixels were encoded as 1, white pixels as 0. Given the vector B of bit values, the complexity C of a given letter i was defined as the summed difference in pixels between that letter and the other letters $j_{1,\ldots,26}$:

$$C_i = \sum_{j=1}^{26} \sum_{k=1}^{400} |B_{ik} - B_{jk}| \tag{2.1}$$

where $k_{1,2,\ldots,400}$ are the indexes of the pixels.

Equation 2.1 quantifies the prototypicality of the visual features of a letter. Values of C_i are low for letters that are similar to many other letters, such as o and c, and high for letters that are dissimilar to most other letters, such as y or w. To obtain the complexity of the visual input for a word w we summed over the complexities C of the letters i:

$$\text{Complexity}_w = \sum_{i=1}^{n} C_i \tag{2.2}$$

where n is the number of letters in the word.

The *Complexity* measure is an obvious simplification of the complex processes involved in the uptake of visual information and merely serves as an approximation of the processing costs associated with the decomposition of a visual word form into orthographic features. Given that the uptake of visual information is not part of the linguistic core of the model this approximation suffices for the current purposes. In the discussion section of this chapter we will briefly discuss alternative implementations of a visual input interpretation mechanism.

2.3.3 Orthography to lexemes

The first part of the linguistic core of the NDR_a model consists of a Rescorla-Wagner network that maps orthographic units onto lexical representations. The orthographic input cues in this network are letters and letter bigrams. For instance, for the word *bear* the input units are the letters *b*, *e*, *a*, *r* and the letter bigrams *#b*, *be*, *ea*, *ar* and *r#*. Richer encodings could be used, but the one adopted here is simple and proved sufficient for the present purposes.

The outcomes of the orthography to lexeme learning network are lexical representations. For the word *bear*, for instance, the outcome is the lexeme *BEAR*. In addition to the lexeme of the target word, we allowed the orthographic input units to co-activate the lexemes corresponding to other words. The orthographic word form *bear*, for instance, co-activates the lexemes *YEAR* and *FEAR*. The co-activation of orthographic neighbors predicts neighborhood and consistency/regularity effects and allows for lexical route processing of non-words. The number of co-activated words taken into consideration is a technical parameter of the model. In all simulations reported in this study this parameter was set to 20. Simulation accuracy is highly similar across a wide range of parameter settings and asymptotes for higher values.

The activation of a word's lexeme given its orthographic representation is defined as the sum over the weights from the letters and letter bigram cues to the lexeme outcome (see Equation 1.3) and will henceforth be referred to as *ActLexeme*. All activations were transformed to positive values by adding the absolute value of the minimum activation.

Furthermore, we back off from zero by adding a small back-off parameter (b, set to 0.10) to the activations. This prevents division by zero when we generate simulated naming latencies.

2.3.4 Lexemes to phonology

The mapping from lexical representations to phonology occurs through a similar Rescorla-Wagner learning network. This network maps lexical representations onto phonological units. As before, lexical representations are lexemes. The phonological units are demi-syllables (Klatt, 1979). The target word *bear*, for instance, consists of two demi-syllables: *b8* and *8R* (using the DISC notation from the CELEX lexical database, see Baayen et al., 1995).[3] Again, it is important to note that demi-syllables are merely a practically convenient approximation of the acoustic gestures necessary for speech production. We return to this issue in the discussion section of this chapter.

While the activation flow in the NDR_a is from lexemes to demi-syllables, we trained the model with demi-syllables as input cues and lexemes as outcomes. This training regime optimizes discriminative learning, because it uses a one-to-many rather than a many-to-one mapping (Ramscar et al., 2010).

The activation of a demi-syllable is obtained by summing over the weights on incoming connections from the active lexemes. The majority of activation spreads from the target word lexemes. We refer to this activation from the target lexeme as a_t. Additional activation $a_{1,\ldots,n}$ spreads to a target demi-syllable from the lexical representations of orthographic neighbors. Given the orthographic input *bear*, for instance, the activations of lexemes of the orthographic neighbors $YEAR$ and $FEAR$ are 0.044 and 0.027. We weighted the contribution of co-activated lexemes to a demi-syllable for the amount of activation they received from the target

[3] Note that these representations are approximations of the demi-syllables used in the speech recognition literature. In our demi-syllables vowels are repeated, whereas in acoustic applications they are split at maximum intensity.

word orthography (w_i). Thus, the activation of a demi-syllable k is given by:

$$\text{ActPhon}_k = w_{lex} * a_t + \sum_{i=1}^{n} w_i * a_i \tag{2.3}$$

where n is the number of lexical neighbors taken into account (set to 20 in the current simulations). In the current simulations, co-activated neighbors were selected from a restricted set of mono-syllabic and mono-morphemic words that can be used as nouns (for more information, see the Simulations section). As before, activations were transformed to positive values by adding the absolute value of the minimum activation and a small back-off parameter (0.10) was added to all activations to prevent division by zero when generating simulated naming latencies.

The parameter w_{lex} indicates the relative weight of the activation from the target lexeme as compared to the activation from lexical neighbors and was set to 4.20 in the current simulations. As such, the activation of a demi-syllable from the target lexeme has a greater weight than the activation from the lexemes of co-activated orthographic neighbors. This is possible only if the language processing system is able to verify that the target lexeme corresponds to the orthographic input, whereas the lexemes of co-activated neighbors do not. Importantly, this assumption is not unique to the NDR_a. Instead, it is a general assumption of discrimination learning that is necessary to evaluate if the outcome of a learning event is predicted correctly and, consequently, to update the association strengths between the cues that are present in the input and all outcomes.

The fact that the NDR_a performs optimally when the relative weight of the activation from the target lexeme is greater than the relative weight of the activation from the lexical neighbors suggests that while lexical neighbors spread activation to demi-syllables during initial bottom-up processing, this activation is suppressed during subsequent processing stages due to top-down verification of the activated lexemes vis-a-vis the current orthographic input. As such, the architecture of the NDR_a is consistent with the idea that successful processing may be characterized

by a bi-directional pass of information between higher and lower level cortical representations (Friston, 2005). As we demonstrate below, the bottom-up pass of information through the principles of discrimination learning captures a wide range of effects observed in naming latencies. The principles underlying the verification processes in the backward top-down information pass, by contrast, are much less well-understood. We return to this issue when discussing the pronunciation performance of the NDR_a model.

Two demi-syllables need to be activated for the mono-syllabic words in this study. We refer to the activation of these demi-syllables as $ActPhon_1$ and $ActPhon_2$. The activation of two demi-syllables introduces a choice problem: one of the activated demi-syllables has to be articulated first. The more dissimilar the activations of the demi-syllables, the harder it may be to produce the right demi-syllable at the right time. A relatively high activation of the second demi-syllables, for instance, may interfere with the production of the first demi-syllable. We model the difficulty of the selection of the appropriate demi-syllable by taking the Shannon entropy (Shannon, 1948) over the activations (transformed into probabilities p_1 and p_2) of the first and second demi-syllable. We refer to this measure as H, which is defined as:

$$\begin{aligned} p_1 &= \text{ActPhon}_1/(\text{ActPhon}_1 + \text{ActPhon}_2), \\ p_2 &= \text{ActPhon}_2/(\text{ActPhon}_1 + \text{ActPhon}_2), \\ \text{H} &= -\sum_{i=1}^{2}(p_i * \log_2(p_i)). \end{aligned} \tag{2.4}$$

2.3.5 Simulating naming latencies

Together, the measures $Complexity$, $ActLexeme$, $ActPhon_1$, $ActPhon_2$ and H describe the total amount of bottom up support for the target

pronunciation.[4] Simulated naming latencies in the NDR$_a$ are modeled through a multiplicative integration of these measures:

$$RT \quad \propto \quad \frac{\text{Complexity}^{w_1}}{\text{ActLexeme}^{w_2} * \text{ActPhon}_1^{w_3} * \text{ActPhon}_2^{w_4} * \text{H}^{w_5}}. \tag{2.5}$$

where $w_{1,\ldots,5}$ are weight parameters that establish the relative contribution of each source of information.

Model parameters were chosen to optimize the quantitative and qualitative performance of the model. For the current simulations, we used the following parameter settings: $w_1 = 2.26$, $w_2 = 0.48$, $w_3 = 0.43$, w_4: 0.49 and w_5: 1.07. Parameter settings were identical in all simulations reported in this study. Including the two technical parameters described earlier (i.e., the back-off parameter that prevents division by zero (0.10) and the number of co-activated neighbors taken into consideration (20)), as well as the parameter for the relative importance of demi-syllable activations from the target word lexeme and the lexemes of lexical neighbors (4.20), the NDR$_a$ has a total of 8 free parameters.

2.4 Simulations

2.4.1 Training and test data

For all simulations described below we trained the orthography-to-lexeme network of the NDR$_a$ on the input lexicon described by Baayen et al. (2011). This training set contains a large number of two and three word phrases from the British National Corpus (Burnard, 1995) that consist of words in a precompiled list of nouns, verbs, adjectives and function words in the CELEX lexical database (Baayen et al., 1995). The lexeme-to-phonology network was trained on 3,908 lowercase mono-morphemic

[4] Lexeme activations for non-words were, by definition, not available. For non-words, *ActLexeme* was therefore set to 0.10 (0 plus the back-off constant b that was added to all lexeme activations for words).

mono-syllabic words in CELEX that consisted of at least 3 letters[5] and for which frequency counts were available in the English Lexicon Project (henceforth ELP, Balota et al., 2004).

For the word naming simulations, we used a data set consisting of the 2,524 mono-morphemic mono-syllabic words present in our training data that can be used as nouns and for which naming latencies are available in the ELP. Prior to analysis we inverse transformed (-1000/RT) the observed naming latencies to remove a rightward skew from the distribution of latencies. In addition, to allow for a comparison of effect sizes, we standardized observed and simulated latencies by converting them to z-scores.

No large-scale database of naming latencies is available for non-words. We therefore extracted a set of non-words from the ARC non-word database (Rastle et al., 2002). We restricted the range for non-word length to that observed in our set of real words and extracted non-words with orthographically existing onsets and bodies only. Furthermore, we restricted the non-words to the words for which both demi-syllables existed in our training lexicon. This resulted in a non-word data set that consisted of 1,822 non-words: 912 regular non-words and 910 pseudo-homophones.

We looked at the effects of 16 linguistic predictors, related to the length, frequency, neighborhood characteristics, regularity/consistency, morphology and semantics of a word or non-word. Predictor values were extracted from the ELP and the *english* data set in the *languageR* package (Baayen, 2011b). Whenever necessary, a more detailed description of each predictor will be provided prior to the description of the results for that predictor.

[5] Although it is not inconceivable that full-form orthographic representations exist for very short words, we decided to follow Baayen et al. (2011) in excluding 1 and 2 letter words (1.48 % of all monosyllabic word types in the CELEX lexical database) from the simulations reported below to prevent biasing the results in favor of a coding scheme that adopts bigram representations at the orthographic level, such as the one used here.

2.4.2 Model evaluation

Model evaluation in cognitive psychology typically involves comparing a model's performance to both observed naming latencies and alternative model architectures. The observed data used in our simulations are the ELP naming latencies for the set of 2524 mono-morphemic nouns described above. We compare the NDR_a model not only to the observed naming latencies, but also the dual-route CDP+ model, which, as noted above, represents the current state of the art in dual route models. Perry et al. (2007) showed that the CDP+ model drastically outperforms other existing models, such as the CDP, the DRC and the triangle model.

The successor of the CDP+ model for bi-syllabic words, the CDP++ model (Perry et al., 2010), has the same architecture for mono-syllabic word reading as the CDP+ model. Due to small changes in parameter settings, the CDP++ shows a small improvement over the CDP+ model in terms of item-level correlations. This, however, comes at the cost of failing to simulate the effect of body neighborhood density. We chose to compare the NDR_a to the CDP+ rather than the CDP++ model, because the former was specifically designed for monosyllabic word reading and its performance in this domain is better documented than that of the CDP++ model. We therefore simulated naming latencies for both the NDR_a and CDP+ model for the set of $2,524$ mono-morphemic nouns under investigation.

2.4.3 Simulation approach

The adequacy of a model can be investigated by comparing its predictions against observed data. This comparison typically focuses on two levels of description. The first level is the overall fit of the model to a set of observed data. Typically, this overall fit is gauged through the *regression approach*. In the regression approach item-level correlations between simulated and observed naming latencies are compared for a large-scale database of words. Here, we follow this approach by looking at the item-level correlations between the ELP naming latencies and the latencies simulated by the NDR_a and CDP+ models. We also look at the posterior

probability of the models as gauged through the Aikake Information Criterion (henceforth AIC, Akaike, 1974). To further probe the overall performance of both models we furthermore conduct a regression analysis on the principal components extracted from the multidimensional space described by all predictors in our simulations. This provides more insight into how well each model captures the overall structure in the observed data.

The second level at which the performance of a model can be investigated concerns the effects of individual predictors on observed naming latencies. The approach that is most typically used to do this is the *factorial approach*. In the factorial approach patterns of results related to predictors are simulated on an experiment-by-experiment basis (for an application, see, e.g., M. Coltheart et al., 2001; Perry et al., 2007). As noted by Adelman and Brown (2008), however, there are a number of problems with the factorial approach.

First, the data gathered in single experiments tend to provide an incomplete picture of the effect of a predictor. The experimental data that models of reading aloud are assessed on are often acquired in experiments with a limited number of carefully selected items and under different experimental conditions. As a result, optimizing the parameter set of a model on the basis of individual experiments may lead to local over-fitting. The model then becomes overly sensitive to the potentially idiosyncratic experimental conditions, item lists and predictor combinations in individual experiments, which comes with the cost of a suboptimal overall model fit (see, e.g., Seidenberg & Plaut, 2006).

Second, modeling on an experiment-by-experiment basis makes it hard to compare the relative effect sizes of different predictors. Due to variations in item lists, experimental conditions and participant populations, the effect sizes for a given predictor can vary substantially between experiments. Given this variance in the effect sizes for a *given* predictor, it is hard to compare effects sizes *between* predictors in the factorial approach.

Third, a large number of experiments are based on factorial contrasts. This leads to a potential distortion of non-linear patterns of results that can range from a simplification of a non-linear effect to masking a predictor effect completely. Applying a median-split dichotomization to a predictor that has a U-shaped effect on response latencies, for instance, would yield a null effect.

To overcome these problems with the factorial approach we adopt a different simulation philosophy. Instead of looking at predictor effects on an experiment-by-experiment basis we will investigate the effects of all relevant predictors in the naming latencies for the set of $2,524$ words in the ELP. All of the ELP naming latencies were obtained in the same task, under very similar experimental conditions and for a homogenous participant population. The presence of an effect in the ELP is a clear indication that computational models should account for this effect. In addition, using ELP naming latencies allows for a comparison of effect sizes between predictors. Furthermore, it allows us to look at the effects of different predictors in a setting where parameters should not be allowed to vary. Finally, because we have access to naming latencies for individual items we can get away from the dichotomization of numeric predictors and start investigating non-linear predictor effects.

2.4.4 Predictor simulations

We investigated a large number of effects that have been documented in the experimental reading aloud literature. For each effect under investigation, we first verified whether an effect was present in the ELP naming latencies. For a large majority of the effects documented in the literature this was indeed the case. Whenever an effect was not present in the ELP naming latencies we explicitly mention its absence. For those effects that were present in the ELP naming latencies we proceeded with an analysis of the effect for the simulated latencies of the NDR_a and CDP+ models.

To investigate the effects of predictors we used the implementation of generalized additive models (henceforth GAMs; Hastie & Tibshirani, 1986) provided by the R package *mgcv* (Wood, 2006). GAMs are an extension of generalized linear models that allow for the modeling of

non-linearities. For each predictor effect we fitted both a linear and a non-linear GAM. The linear GAM is mathematically equivalent to a simple linear regression model. This linear model provides a conventional assessment of the presence or absence of predictor effect. In addition it provides an effect size measure that allows for the comparison of the relative magnitude of effects of different predictors.

The non-linear GAMs allow us to capture non-linearities. The smooth functions in GAMs do not presuppose particular non-linear structures and can therefore model a wide range of predictor-related non-linearities. Furthermore, tensor products allow us to model two-dimensional non-linear interactions of numerical predictors. As a result, we do not have to dichotomize predictors even when inspecting interaction effects. We allowed all predictor smooths to describe up to 6th order non-linearities ($k = 6$) and did not impose any restrictions on tensor products. We removed predictor values further than 3 standard deviations from the predictor mean in all non-linear GAMs to prevent smooth estimates from being overly influenced by extreme predictor values. To establish the significance of tensor product interactions, we compared the AIC score (Akaike, 1974) of a tensor product GAM to that of a GAM with additive non-linear effects of both predictors (i.e., separate predictor smooths). Unless explicitly stated otherwise, we considered interactions only when the AIC score of the tensor product GAM was significantly lower than that of a GAM with an additive non-linear effect of both predictors.

Many of the predictors under investigation are strongly correlated. As a result introducing a model term to or removing it from a model that contains all predictors could have a strong effect on the effects of the other terms in the model. To side-step this problem of multicollinearity we decided to fit separate models for each predictor. Fitting separate models for each predictor comes at the cost of masking potential effects of covariates. We addressed this problem in two ways. First, we simulated only effects that have been documented in experimental studies with carefully controlled item lists. Second, to ensure that our model captures the joint effects of the predictors we conducted a principal components

regression analysis for both observed and simulated latencies on the multidimensional input space described by all 16 predictors.

For word naming we fitted models to the observed naming latencies, as well as to the simulated naming latencies for the NDR$_a$ and CDP+ models. As mentioned earlier, no large-scale database of non-word naming latencies exists. To simulate the non-word effects documented in the literature we therefore could not compare our model to observed naming latencies. We did, however, have the possibility of comparing non-word naming performance in the NDR$_a$ and CDP+ models. This allows us to establish whether or not the single-route architecture of the NDR$_a$ model captures the experimental effects of non-word naming that are successfully simulated by the CDP+ model. Furthermore, it allows us to identify whether and where predictions for non-word naming differ between the NDR$_a$ and CDP+ models. These differences describe explicit test-cases for the performance of both models that can be addressed in future non-word naming experiments.

2.5 Simulation results

2.5.1 Non-word naming disadvantage

Before we turn to the discussion of predictor-specific effects, there is an overall difference between word and non-word naming that requires our attention. Several studies have documented that words are named faster than non-words (McCann & Besner, 1987; Weekes, 1997; Ziegler et al., 2001). Both models correctly predict this effect (NDR$_a$: t = -13.090, β = -0.395; CDP+: t = -74.142, β = -1.514). There is, however, a large difference in the relative magnitude of the predicted effects. In the CDP+ model, mean naming latencies for non-words are 57% slower than those for words (159 vs 101 cycles). In the NDR$_a$, the difference is only 26% (mean inverse activation units non-words: 2,275,498, words: 1,809,877).

Although a direct comparison to the effect size in observed data is not possible, we compared the processing disadvantage for non-words predicted by the NDR$_a$ and CDP+ models to that observed in the studies of

McCann and Besner (1987) and Ziegler et al. (2001). The average naming latency for words in McCann and Besner (1987) was 454 ms, whereas that for non-words was 579 ms. The processing disadvantage for non-words in this study was therefore 29%. In Ziegler et al. (2001), average naming latencies across eight conditions were 611 ms for non-words and 521 ms for words, for a non-word processing disadvantage of only 17%. These data suggest that the CDP+ model overestimates the processing costs for non-words, while the NDR_a provides a more reasonable estimate.

2.5.2 Length effects

2.5.2.1 Word length. The effect of word length on naming latencies has been documented in a large number of studies, with longer naming latencies for words that consist of more letters (see, e.g., Richardson, 1976; Frederiksen & Kroll, 1976; Henderson, 1982; Balota & Chumbley, 1985; Jared & Seidenberg, 1990; Seidenberg & McClelland, 1990; Spieler & Balota, 1997; Weekes, 1997). This length effect is present in the ELP naming latencies ($t = 20.047$, $\beta = 0.371$), as well as in the NDR_a ($t = 46.315$, $\beta = 0.678$) and CDP+ simulations ($t = 20.061$, $\beta = 0.371$). The results of a non-linear model are presented in Figure 2.5 and indicate that this effect is slightly non-linear for the observed naming latencies and the latencies simulated by the NDR_a, but not for those of the CDP+ model.

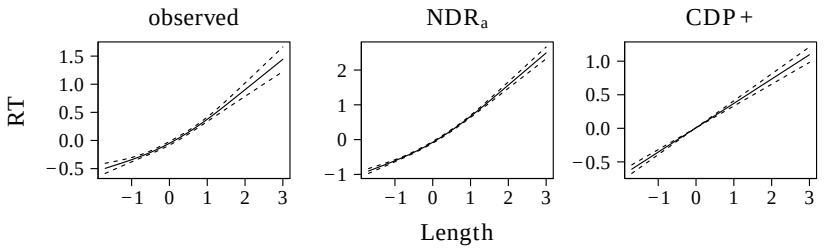

Figure 2.5. The effect of length in word naming.

The effect size of the length effect is larger in the NDR$_a$ model than in the observed data. The length effect in the NDR$_a$ model is primarily driven by the complexity of the visual input. In all reported simulations, the visual complexity parameter is set to 2.08. The overall fit of the data, however, is quite robust to changes in this parameter setting (e.g., overall data fit: $r \geq 0.45$ for parameter values between 0.98 and 3.16). There are two reasons we decided to use the current parameter setting. First, we believe that the overall fit of the model should be optimal. This is the case for the current parameter settings. Second, because the model operates under noise-free conditions, the effect sizes in the NDR$_a$ tend to be somewhat larger than those in the observed data. As we will show in the overall model fit section of this chapter, the effect size of length in the current simulations is of the correct relative magnitude compared to the effect sizes of the other predictors.

In addition to a length effect for words, a length effect for non-words has also been observed. Non-word naming latencies increase linearly for each additional letter (Weekes, 1997; Ziegler et al., 2001). Both the NDR$_a$ ($t = 130.845$, $\beta = 0.951$) and the CDP+ ($t = 21.236$, $\beta = 0.446$) capture the effect of length in non-word naming. As can be seen in Figure 2.6, both models predict a linear (CDP+) or near-linear (NDR$_a$) effect for non-words. Furthermore, consistent with the experimental findings of Weekes (1997) and Ziegler et al. (2001), both models predict a larger effect size for length in non-word naming than in word naming (NDR$_a$: $\Delta\beta = 0.273$, CDP+: $\Delta\beta = 0.075$) . The relative magnitude of the length effect for non-words as compared to that for words is somewhat larger in the NDR$_a$ ($\frac{\beta_{nw}}{\beta_w} = 1.402$) than in the CDP+ ($\frac{\beta_{nw}}{\beta_w} = 1.201$).

In addition to the effects of word length reported above, Weekes (1997) also reported an interaction of length with frequency, with a stronger length effect for low frequency words. In a reanalysis of the Weekes (1997) data, Perry et al. (2007), however, demonstrated that this interaction was not significant. For the current set of observed naming latencies the interaction was not supported either: a model with additive non-linear terms of frequency and length resulted in a lower AIC score than a GAM with a tensor product of frequency and length.

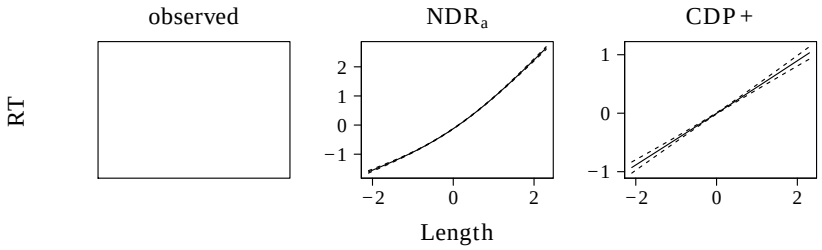

Figure 2.6. The effect of length in non-word naming. The panel for the observed data is left blank, because no large-scale database of non-word naming exists.

2.5.3 Neighborhood effects

2.5.3.1 Orthographic Neighborhood Size. Although the unique variance accounted for by neighborhood measures is small (Baayen et al., 2006), these effects have played a central role in the assessment of models of reading aloud. The experimental naming literature has consistently documented that words with many orthographic neighbors are processed faster than words with fewer neighbors (Andrews, 1989, 1992, 1997; Grainger, 1990; V. Coltheart et al., 1988). In interactive activation models, however, the inhibitory links between lexical items lead to more competition for words with many orthographic neighbors. As a result, the DRC model, which uses the interactive activation model of McClelland and Rumelhart (1981) as its lexical route, could only model the effect of orthographic neighborhood density with altered parameter settings. Although the CDP+ model uses the same interactive activation architecture for its lexical route, it captures the orthographic neighborhood density effect, presumably through its non-lexical route. Nonetheless, the authors acknowledge that the interactive activation model in their lexical route may have inherent problems with neighborhood density effects and that there may be better alternatives for the lexical route of the CDP+ model (Perry et al., 2007, p.303).

The NDR$_a$ model predicts orthographic neighborhood density facilitation as a consequence of the co-activation of orthographically similar words. Each co-activated orthographic neighbor activates its lexeme, from which in turn activation spreads to the corresponding demi-syllables. The target word *band*, for instance, co-activates the lexical representations of words like *bank*, *bang* and *ban*, which spread activation to the target demi-syllable *b{*. In addition, *band* co-activates *land*, *hand* and *sand*, which spread activation to the target demi-syllable *{nd*. The more orthographic neighbors a word has, the more activation will spread from co-activated lexemes to the target demi-syllables and the faster a word will be named.

A linear model on the ELP naming latencies shows the predicted facilitatory effect of orthographic neighborhood density (t = -19.173, β = -0.357). Both the NDR$_a$ (t = -25.080, β = -0.447) and the CDP+ (t = -19.211, β = -0.357) capture this linear effect of orthographic neighborhood density. The non-linear effect of orthographic neighborhood density is shown in Figure 2.7. The NDR$_a$ model predicts a quadratic curve that is highly similar to that in the observed data, whereas CDP+ captures the linear trend of the effect, but somewhat underestimates its quadratic component.

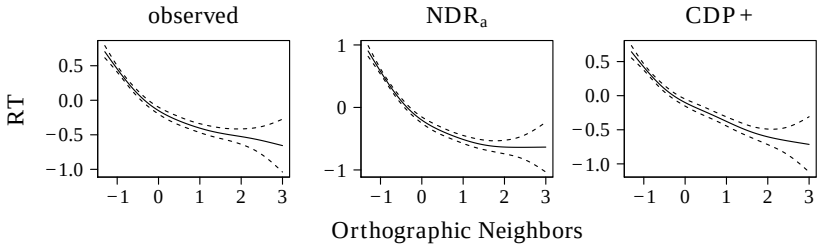

Figure 2.7. The effect of orthographic neighborhood density in word naming.

In addition to an effect of orthographic neighborhood density on word naming, a non-word naming effect has also been documented (see e.g., Andrews, 1997). As for real words, the effect is facilitatory in nature, with faster naming latencies for non-words with many orthographic neighbors

as compared to non-words with frew orthographic neighbors. A linear model on the simulated naming latencies shows that both the NDR$_a$ (t = -27.595, β = -0.160) and the CDP+ model (t = -14.014, β = -14.014) correctly simulate this effect. Furthermore, as can be seen in Figure 2.8, the NDR$_a$ predicts a quadratic non-linearity that is similar to the observed effect of orthographic neighborhood density in word naming. The CDP+ predicts a qualitatively similar, but somewhat more wriggly non-linear effect.

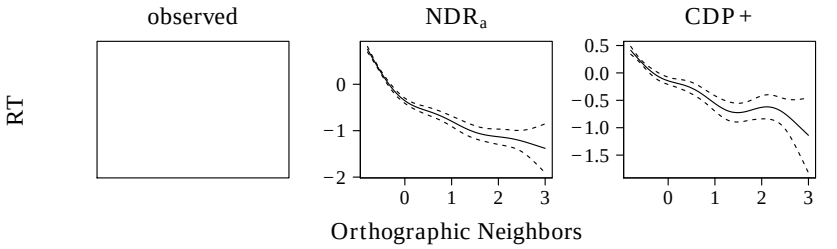

Figure 2.8. The effect of orthographic neighborhood density in non-word naming.

2.5.3.2 Phonological and Body Neighborhood Size. The effect of orthographic neighborhood density is not the only neighborhood density effect that has been documented. As noted by Perry et al. (2007), several studies have argued that phonological neighborhood density (Mulatti et al., 2006) or body neighborhood density (Brown, 1987; Jared et al., 1990; Ziegler et al., 2001) may be more adequate measures of neighborhood density effects in reading aloud. The linear effect of phonological neighborhood density (observed: t = -12.760, β = -0.246; NDR$_a$: t = -16.440, β = -0.311; CDP+: t = -15.332, β = -0.292), as well as that of body neighborhood density (observed: t = -5.751, β = -0.114; NDR$_a$: t = -4.793, β = -0.095; CDP+: t = -13.991, β = -0.268) is captured by both the NDR$_a$ and the CDP+ model. The NDR$_a$ model somewhat underestimates the magnitude of the body neighborhood density effect relative to that of the

orthographic and phonological neighborhood density effects. By contrast, the CDP+ model overestimates the effect of body neighborhood density.

The non-linear effect of phonological and body neighborhood density is presented in Figure 2.9. The top panel of this figure shows that both the NDR_a and the CDP+ correctly simulate the overall quadratic nature of the phonological neighborhood density effect in the ELP, although the simulated effect in the NDR_a is more similar to the observed effect than is the simulated effect in the CDP+ model. The bottom panel of Figure 2.9 shows that the effect of body neighborhood density is u-shaped in nature, with particular difficulties for words with few body neighbors. Both models have trouble capturing the non-linear effect of body neighborhood density, but the deviations of the simulated effect from the observed effect are greater for the NDR_a than for the CDP+ model.

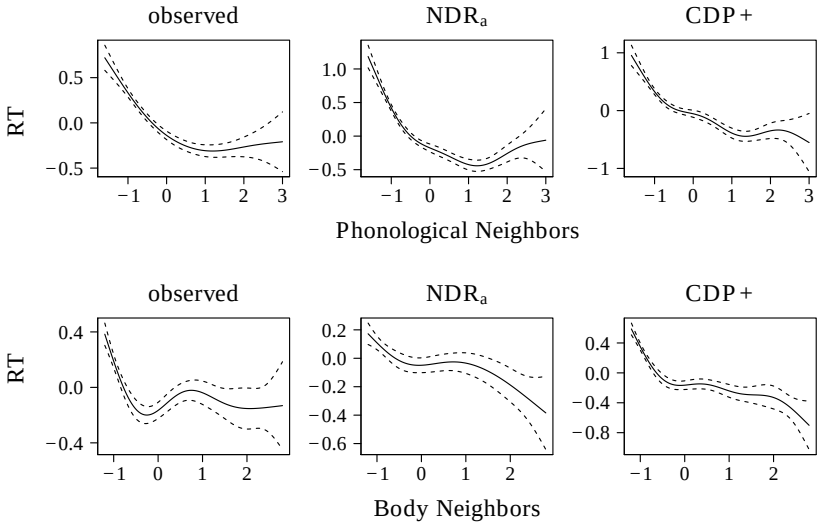

Figure 2.9. The effects of phonological neighborhood density and body neighborhood density in word naming.

In addition to a word naming effect, both the NDR_a and the CDP+ model predict a non-word naming effect of both phonological (NDR_a: t = -16.724, β = -0.034; CDP+: t = -9.238, β = -0.020) and body neigh-

borhood density (NDR$_a$: t = -7.560, β = -0.032; CDP+: t = -9.168, β = -0.039). As for the effect of orthographic neighborhood density, the non-linear estimates of these effects are qualitatively similar in both models, although the quadratic component of the phonological neighborhood density effect is more pronounced in the NDR$_a$, whereas the quadratic component of the body neighborhood density effect is more pronounced in the CDP+ (see Figure 2.10).

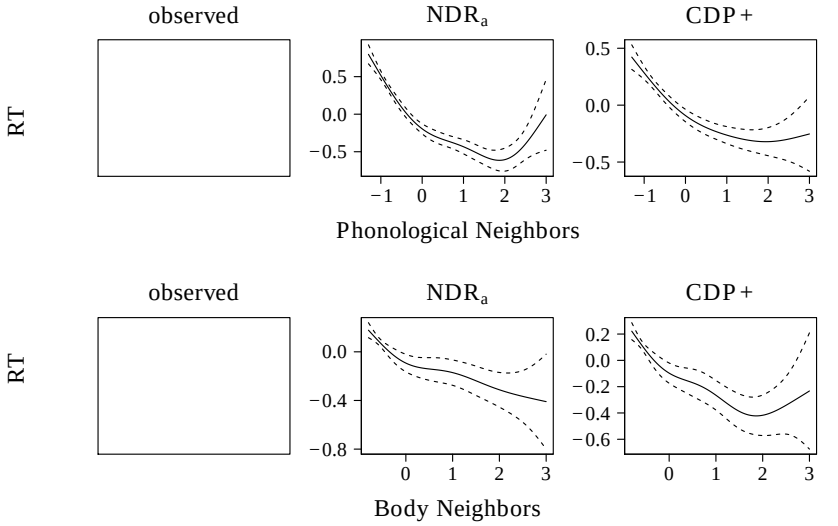

Figure 2.10. The effects of phonological neighborhood density and body neighborhood density in non-word naming.

2.5.3.3 The interplay of neighborhood density measures. As noted above, the NDR$_a$ predicts that the effect of neighborhood density is primarily an orthographic neighborhood density effect, whereas several studies have argued that phonological or body neighborhood density characteristics may underlie the effect of orthographic neighborhood density. To investigate which neighborhood density measure drives the neighborhood effects, we entered all three predictors into a single linear regression model. Table 2.1 shows t-values and β coefficients for the neighborhood

Table 2.1. The linear interplay of orthographic, phonological and body neighborhood density. Listed are t-values and β coefficients for each of the predictors in an additive linear model

	observed		NDR$_a$		CDP+	
	t	β	t	β	t	β
Orthographic N	-13.859	-0.368	-19.389	-0.490	-8.151	-0.214
Phonological N	-0.948	-0.023	-1.170	-0.027	-5.059	-0.122
Body N	2.790	0.058	6.690	0.133	-6.942	-0.144

density measures in this model. When taking the effect of orthographic neighborhood density into account, phonological neighborhood density no longer has a significant effect on the observed naming latencies and body neighborhood density shows a small inhibitory effect, which may be due to suppression (L. Friedman & Wall, 2005).

The NDR$_a$ model captures the general pattern of results: orthographic neighborhood density remains highly significant, the effect of phonological neighborhood density disappears and body neighborhood density becomes inhibitory. As in the individual models for the three predictors, however, the NDR$_a$ somewhat underestimates the effect of body neighborhood density, which is reflected in an overly large *positive* t-value. The CDP+ model has more problems with the interplay of the neighborhood density predictors. It underestimates the contribution of orthographic neighborhood density and incorrectly predicts strong inhibitory effects for both body and phonological neighborhood density.

To further explore the interplay of the neighborhood density measures, we fitted two GAMs to assess the potential non-linear interplay of orthographic neighborhood density with phonological and body neighborhood density. The first GAM included a tensor product of orthographic neighborhood density and phonological neighborhood density, the second a tensor product of orthographic neighborhood density and body neighborhood density. Both models provide a better account of the data than models with separate smooths for both predictors, as indicated by lower AIC scores.

The results of the tensor product models are shown in Figure 2.11. In the observed naming latencies a strong facilitatory effect of orthographic neighborhood density characterizes both models. Both phonological and body neighborhood density show an effect only for words with many orthographic neighbors. For these words, the effect of phonological neighborhood density is inhibitory, whereas that of body neighborhood density is facilitatory. The NDR$_a$ model correctly simulates both numerical interactions and shows a pattern of results that is highly similar to that in the observed data. The simulations of the CDP+ model show more deviation from the observed data. For words with many orthographic neighbors the CDP+ model incorrectly predicts a facilitatory effect of phonological neighborhood density. Furthermore, the CDP+ model underestimates the effect of orthographic neighborhood density for words with few body neighbors or few phonological neighbors.

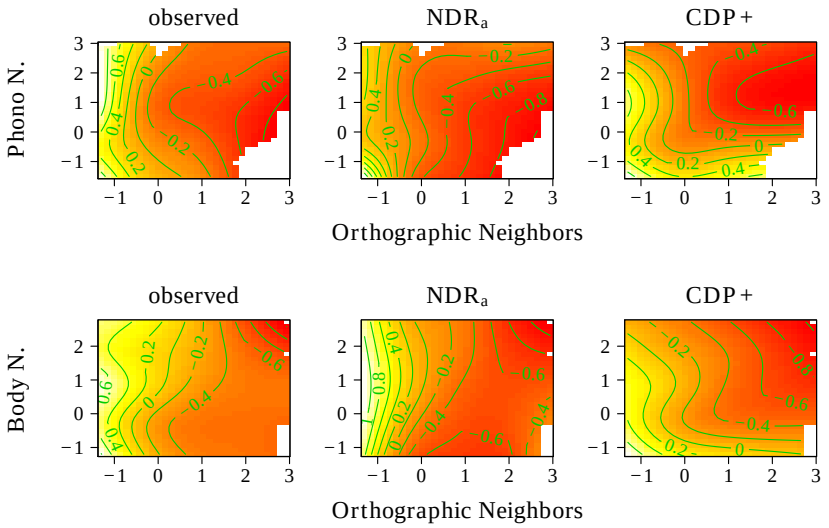

Figure 2.11. The interplay of orthographic, phonological and body neighborhood density in tensor product GAMs.

Two conclusions can be drawn from the results of these simulations. First, the neighborhood density effect seems to primarily be an effect of *orthographic* neighborhood density. This argues against an interpretation of neighborhood effects as being driven by phonological or body neighborhood density. Second, the tensor product GAMs on the observed data show that the effect of orthographic neighborhood density is modulated by phonological and body neighborhood density for words with many orthographic neighbors. This effect is facilitatory for body neighborhood density and inhibitory for phonological neighborhood density. The correct characterization of this pattern by the NDR_a suggests that the model is sensitive to the neighborhood similarity structure that characterizes the English lexical space.

It is worth taking a moment to consider why the NDR_a model captures the complex interplay of the neighborhood density measures. Neighborhood effects in the NDR_a arise due to bottom-up co-activation of orthographic neighbors of the target word. When the orthographic word *bear* is presented, for instance, not only the corresponding lexeme *BEAR* is activated, but lexemes of orthographic neighbors such as *PEAR*, *WEAR*, *HEAR* and *YEAR* receive activation as well. The more lexemes are co-activated, the more activation spreads from these co-activated lexemes to the phonological level. The fact that the neighborhood density effects seem to primarily be driven by orthographic neighborhood density therefore follows straightforwardly from the architecture of the NDR_a model.

From an orthographic perspective, all co-activated lexical representations are equal. In the context of the reading aloud task, however, some co-activated lexical representations are more equal than others. When the word *bear* is presented, the co-activated lexeme *HEAR* does not share a phonological demi-syllable with the target word lexeme *BEAR*. The fact that *HEAR* is co-activated by the orthographic presentation of *bear*, therefore, does not help activate the phonology the target word *bear* faster. By contrast, *PEAR* and *WEAR* share the phonological rhyme with *BEAR* and therefore help reduce the time it takes to activate the second demi-syllable *8R* of the target word *bear*. This explains the

facilitatory effect of body neighborhood density for words with many orthographic neighbors. Body neighbors are words that share both the orthographic and the phonological rhyme with the target word. The more of the orthographic neighbors are body neighbors, the faster the second demi-syllable of the target word is activated and the faster that target word is named.

The effect of phonological neighborhood density is opposite to that of body neighborhood density. For words with many orthographic neighbors the observed naming latencies show an inhibitory effect of phonological neighborhood density: words with many phonological neighbors are named slower than words with few phonological neighbors. As counter-intuitive as this inhibitory effect of phonological neighborhood density might seem, it follows straightforwardly from the architecture of the NDR_a model. In contrast to body neighbors, the lexemes of phonological neighbors are not necessarily co-activated by the orthographic presentation of the target word. The orthographic presentation of the word *bear*, for instance, does not co-activate the lexical representations *HAIR* and *AIR*. *HAIR* and *AIR* therefore cannot help activate the target word phonology, despite the fact that these lexemes share the second demi-syllable with *BEAR*. The model, however, has learned to associate *HAIR* and *AIR* with the word-final demi-syllable *8R*. The higher the number of lexemes that share a demi-syllable, the less well the association between each lexeme and that demi-syllable will be learned. The existence of the phonological neighbors *HAIR* and *AIR* therefore lead to a lower connection strength from the lexeme *BEAR* to the demi-syllable *8R*. This results in a longer naming latency for the word *bear* than would be the case if no such phonological neighbors existed.

2.5.3.4 Pseudo-homophones. As noted by M. Coltheart et al. (2001), the neighborhood density effects reported above are complemented by a pseudo-homophone effect in non-word naming (McCann & Besner, 1987; Taft & Russell, 1992; Seidenberg et al., 1996). Naming latencies for non-words that can be pronounced as real words (e.g., *bloo*) are shorter as compared to naming latencies for normal non-words. This

pseudo-homophone effect is correctly predicted by both models, albeit with relatively small effect sizes (NDR$_a$: t = -4.348, β = -0.203; CDP+: t = -2.679, β = -0.125). In addition, there has been some debate as to whether or not there is a base word (e.g., *blue*) frequency effect for pseudo-homophones. In a review of the evidence, Reynolds and Besner (2005, p.623) conclude that "the published data are most consistent with the conclusion that there is no base word frequency effect on reading aloud when pseudohomophones are randomly mixed with control nonwords". In pure non-word blocks, however, an effect of base word frequency has been observed (Borowsky et al., 2002; Marmurek & Kwantes, 1996). In our non-word simulations, both the NDR$_a$ (t = -3.630, β = -0.119) and the CDP+ (t = -2.668, β = -0.092) yielded a subtle base word frequency effect. The limited size of the predicted base word frequency effect in our large-scale simulations, however, suggests that the effect may be hard to detect in single experiment studies.

2.5.3.5 Orthographic Neighborhood by Frequency. A further important neighborhood density effect concerns the interaction of orthographic neighborhood density with frequency. Several studies found that low frequency, but not high frequency words are read faster when they have many neighbors (Andrews, 1989, 1992; Balota et al., 2004). We fitted tensor product GAMs to look at the interaction of frequency and neighborhood density in the observed and simulated naming latencies. The results

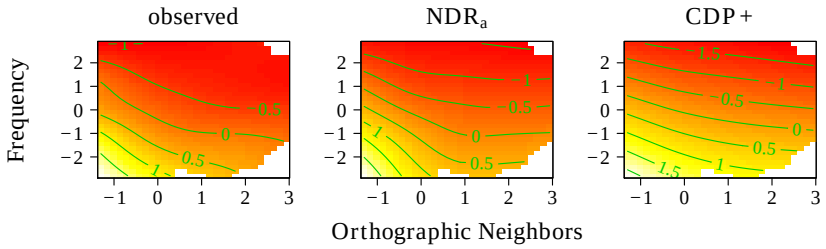

Figure 2.12. The interaction of frequency with orthographic neighbors and phonological neighbors in tensor product GAMs.

of this model are shown in Figure 2.12. The observed data show the expected pattern of results: a facilitatory effect of neighborhood density for low frequency words only. Both the NDR_a and the CDP+ model capture this frequency by orthographic neighborhood density interaction and show the longest latencies for low frequency words with few orthographic neighbors.

2.5.4 Consistency/Regularity effects

2.5.4.1 Regularity. The relation between the orthography and phonology of a word has been a hotly debated topic in the word naming literature. M. Coltheart et al. (2001) focused on the concept of regularity and defined a word as regular "if its pronunciation is correctly generated by a set of grapheme to phoneme conversion rules" (M. Coltheart et al., 2001, p.231). The DRC model predicted that regular words should be pronounced faster than irregular words. This was confirmed by a number of experimental findings (Seidenberg et al., 1994; Taraban & McClelland, 1987; Paap et al., 1987; Paap & Noel, 1991). We therefore consider the effect of regularity a good starting point for the investigation of the relation between orthography and phonology. In our simulations we defined regularity as a two-level factor, based on the regularity of a word given the grapheme to phoneme (henceforth GPC) rules underlying the sub-lexical route of the DRC model. A linear model on the ELP naming latencies shows the predicted facilitation for regular words (t = -8.864, β = -0.389). This effect is somewhat underestimated by the NDR_a (t = -5.762, β = -0.255) and somewhat overestimated by the CDP+ (t = -14.578, β = -0.624).

2.5.4.2 Position of irregularity. The size of the regularity effect depends on the position at which the irregularity occurs. A number of studies (M. Coltheart & Rastle, 1994; Rastle & Coltheart, 1999; Roberts et al., 2003) found larger irregularity effects for words with early-position irregularities as compared to words with late-position irregularities. A similar effect of position of irregularity is present in the ELP naming latencies (t = -5.043, β = -0.746) and the CDP+ simulation (t = -3.039,

β = -0.546). The NDR$_a$ model, however, failed to capture this effect (t = 1.192, β = 0.177). The inability of the NDR$_a$ to model the position of irregularity effect is not surprising given that the model is insensitive to the sequential nature of the orthographic input and the phonological output. We return to this issue in the discussion section of this chapter.

2.5.4.3 Consistency. In a number of studies, it has been argued that regularity may be a measure that is too simplistic to fully describe the relationship between the orthography and the phonology of a word. Following Glushko (1979), a number of studies therefore investigated the effect of consistency, rather than regularity. Originally, Glushko (1979) defined consistency as a two-level factor, for which words were defined as inconsistent if their orthographic body mapped on to more than one phonemic sequence. For instance, while the pronunciation of the word *wave* is correctly predicted by the GPC rules of the DRC model it is inconsistent, because its word body is pronounced differently in the word *have*. Further research indicated that consistency is better conceptualized as a continuous variable (Jared et al., 1990; Plaut et al., 1996; Jared, 1997; Rastle & Coltheart, 1999).

We tested a number of consistency measures and found the proportion of consistent word tokens to explain most variance in the ELP naming latencies (t = -8.212, β = -0.169). This linear effect of consistency was accurately captured by the NDR$_a$ (t = -7.354, β = -0.150), as well as

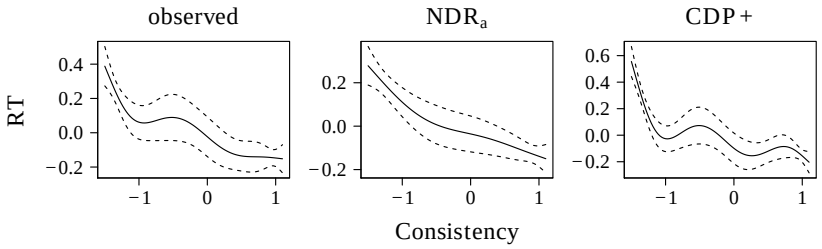

Figure 2.13. The effect of consistency of the orthography to phonology mapping in word naming.

the CDP+ model (t = -9.685, β = -0.188). Figure 2.13 shows the non-linear effect of consistency. The consistency effect was stronger for low predictor values in the observed naming latencies. The NDR$_a$ somewhat underestimates the non-linear component of this effect, whereas the CDP+ somewhat overestimates it. Given the width of the confidence interval for the observed effect of consistency, however, it is unclear how pronounced the non-linearity of the consistency effect in the observed data is.

In addition to a word naming effect of consistency, a consistency effect has also been observed in non-word naming (Glushko, 1979; Andrews & Scarratt, 1998) For our set of non-words, both the NDR$_a$ (t = -8.049, β = -0.185) and the CDP+ (t = -12.607, β = -0.283) predict a facilitatory effect of consistency. As can be seen in Figure 2.14, this effect resembles the non-linear effect in the observed data for the NDR$_a$, whereas the CDP+ shows a more uniform facilitatory effect. Given the width of the confidence intervals for the non-word effect in both models, however, this difference is not statistically robust.

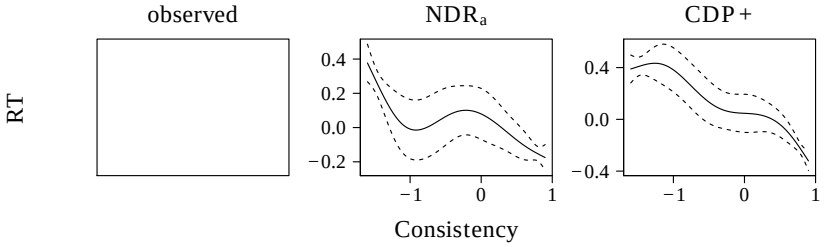

Figure 2.14. The effect of consistency on non-word naming.

Both models predict a larger magnitude of the consistency effect in non-word naming than in word naming, with a somewhat greater difference in the relative magnitudes in the CDP+ model ($\frac{\beta_{nw}}{\beta_w}$ = 1.507) than in the NDR$_a$ ($\frac{\beta_{nw}}{\beta_w}$ = 1.235). The prediction that the consistency effect should be larger in non-word naming as compared to word-naming stands in contrast with the findings in Glushko (1979), who found a 29 ms facilitatory effect for consistent words in both word and non-word naming. In the absence of a large-scale database of non-word naming latencies

or further experimental findings, however, any conclusions regarding the simulation of the relative effect sizes of the effects of consistency in word and non-word naming in the NDR$_a$ and CDP+ models are tentative.

2.5.4.4 Consistency by Regularity. Now that we established the existence of both a consistency and regularity effect, we return to the question of which measure best characterizes the effect of the orthography to phonology mapping on naming latencies. It would be problematic for the DRC model if an independent graded consistency effect were present on top of the regularity effect, because its non-lexical route is based on hard-coded rules that operate in an all-or-none fashion (Andrews & Scarratt, 1998; Zevin & Seidenberg, 2006). In contrast, the CDP+ model is sensitive to the probabilistic characteristics of orthography to phonology mappings (Zorzi et al., 1998b; Zorzi, 1999). This model therefore allows for the possibility that graded consistency might be a better measure than regularity.

In the NDR$_a$ model, regularity and consistency effects originate from the co-activation of lexical items with similar orthographies. The word *band* co-activates the lexical representations of phonologically consistent words like *bank*, *bang* and *ban*. These words provide additional support for the target demi-syllable {*nd* and hence speed up naming latencies. In contrast, *bough* co-actives the lexemes of phonologically inconsistent neighbors, such as *tough*, *rough* and *cough*. The lexemes corresponding to these inconsistent neighbors activate the non-target demi-syllable *Vf* and therefore do not facilitate the pronunciation of the target word *bough*. The amount of support for the target demi-syllables directly depends on the number of co-activated lexemes of orthographically consistent and inconsistent words. The NDR$_a$ therefore explicitly predicts that graded consistency should be a better measure of orthography to phonology mapping effects than regularity.

Inspection of the naming latencies in the ELP revealed not only an independent contribution of both regularity and consistency, but also a significant interaction between regularity and consistency. Table 2.2 shows the results of a linear model that includes regularity, consistency

Table 2.2. The interplay of regularity and consistency. Listed are t-values and β coefficients for each of the predictors in an additive linear model

	observed		NDR_a		CDP+	
	t	β	t	β	t	β
Consistency	-6.174	-0.253	-5.846	-0.239	-5.940	-0.227
Regularity (factor)	-2.982	-0.153	-0.499	-0.026	-6.994	-0.336
Interaction	3.016	0.144	2.739	0.131	2.441	0.109

and a regularity by consistency interaction term. For the observed naming latencies, the strongest effect is that of consistency. The effect of regularity becomes weaker in a model that includes consistency, but remains significant. Furthermore, there is a significant interaction of regularity with consistency. The NDR_a captures the main effect of consistency and the interaction term, but fails to capture the independent contribution of regularity. In the CDP+ model all three effects are significant. In contrast to the NDR_a, however, it overestimates the independent contribution of regularity.

Figure 2.15 shows the non-linear interaction of consistency with regularity, which sheds further light on the issue. Regular words (top row) show a subtle linear effect in the observed data, as well as in the simulations of the NDR_a and CDP+ models. For irregular words, however, a non-linear curve characterizes the ELP naming latencies, with particularly long reaction times for inconsistent irregulars. The general shape of this curve is captured fairly well by both models, although the NDR_a reduces the third-order effect to a second-order curve and the CDP+ overestimates the processing difficulties for inconsistent irregulars.

2.5.4.5 Consistency by Friends-Enemies. Consistency has also been shown to interact with the number of friends (words with the same body and rime pronunciation) and enemies (words with a different body and rime pronunciation) a word has. Jared (1997, 2002), for instance, found an effect of consistency that was limited to words with more enemies than friends. Different friend-enemy measures have been proposed. Here, we use the measure that explained most of the variance in the ELP naming

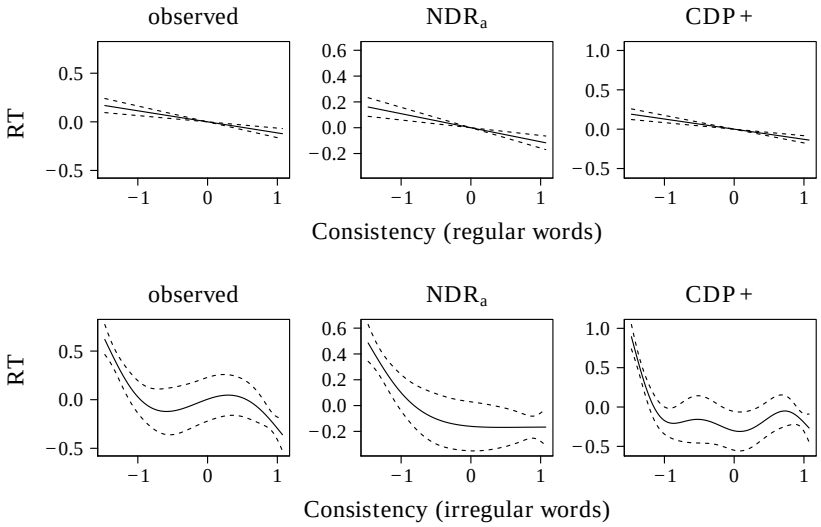

Figure 2.15. The interplay of consistency and regularity in word naming. Top row shows results for regular words, bottom row for irregular words.

latencies, which is the number of friends minus the number of enemies (t = -6.160, β = -0.128). Both the NDR$_a$ (t = -3.779, β = -0.078) and the CDP+ (t = -8.715, β = -0.170) showed a significant main effect of this friend-enemy measure on the simulated naming latencies, although the NDR$_a$ somewhat underestimates its effect size.

More interestingly, the observed data support a tensor product GAM with an interaction between consistency and our friend-enemy measure. This interaction is displayed in Figure 2.16. Consistent with the findings by Jared (2002), the consistency effect in the observed data is stronger for words with more enemies. As can be seen in the middle and right panels of Figure 2.16, both the NDR$_a$ and the CDP+ capture the complex nature of this interaction.

2.5.4.6 Consistency by Frequency. A final effect of consistency/regularity that warrants some discussion is the interaction of frequency with these measures. Jared (1997, 2002) did not find evidence for an interaction of

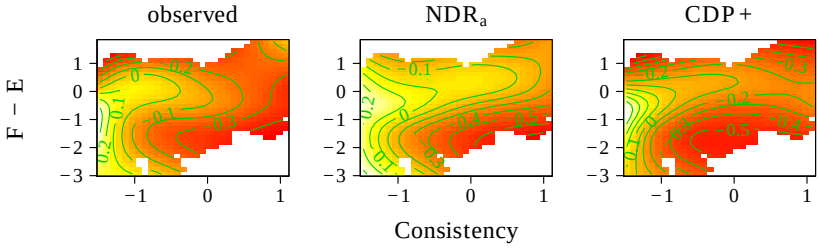

Figure 2.16. The interaction of consistency with friends minus enemies in tensor product GAMs.

either regularity or consistency with frequency. As noted by Perry et al. (2007), these null results stand in contrast to previous studies (Seidenberg et al., 1994; Taraban & McClelland, 1987; Paap et al., 1987; Paap & Noel, 1991) that reported longer naming latencies for irregular or inconsistent low-frequency words, but not for high-frequency words. The ELP naming latencies showed very similar AIC scores for a model with a tensor product interaction of consistency and frequency (AIC: 5015.91) and a model with separate smooths for consistency and frequency (AIC: 5015.59). The evidence for a consistency by frequency interaction in the ELP naming latencies, therefore, is subtle at best. For completeness, we nonetheless show the results of the tensor product GAM in Figure 2.17. The panel for the observed data shows a subtle interaction in the expected direction,

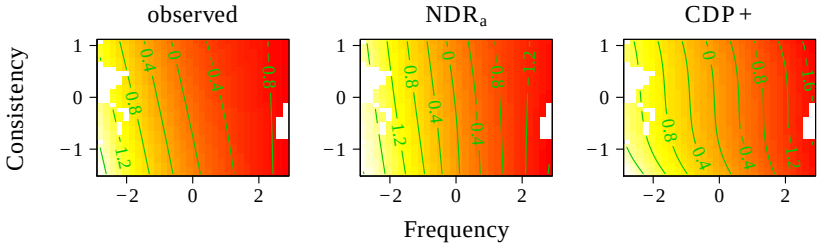

Figure 2.17. The interaction of frequency with consistency in tensor product GAMs.

with a consistency effect that is more prominent for low frequency than for high frequency words. Both the NDR$_a$ and CDP+ simulations show a qualitatively similar subtle interaction. The current simulations therefore suggest that both models are capable of explaining the subtle interplay between consistency and frequency.

2.5.5 Frequency effects

2.5.5.1 Frequency. We have not yet discussed the main effect of the predictor that correlates most strongly with observed naming latencies: word frequency. The effect of frequency is the most well-established effect in the word naming literature (see, e.g., Forster & Chambers, 1973; Balota & Chumbley, 1985; Weekes, 1997; Jared, 2002) and is highly significant in the observed naming latencies (t = -25.523, β = -0.453). As expected, both models capture the frequency effect (NDR$_a$: t = -40.202, β = -0.625; CDP+: t = -40.457, β = -0.627). As can be seen in Figure 2.18 the effect is linear or near-linear in the observed data, as well as in the simulations for both models.

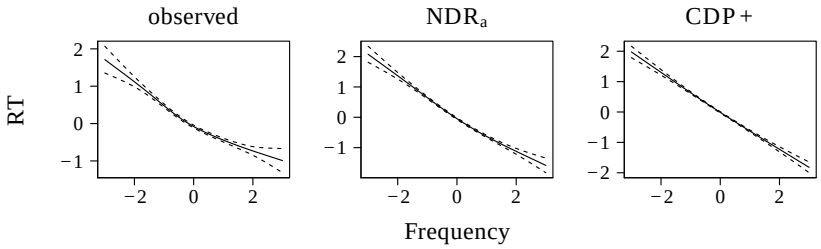

Figure 2.18. The effect of frequency in word naming.

2.5.5.2 Familiarity. In addition to the frequency effect, we also investigated the effect of familiarity on the ELP naming latencies. As can be seen in Figure 2.19, the effect of familiarity in the observed data is linear and highly similar to that of frequency (t = -17.893, β = -0.347). Both models capture the general linear trend (NDR$_a$: t = -23.098, β = -0.424; CDP+: t = -29.091, β = -0.484). While the observed data show a slightly

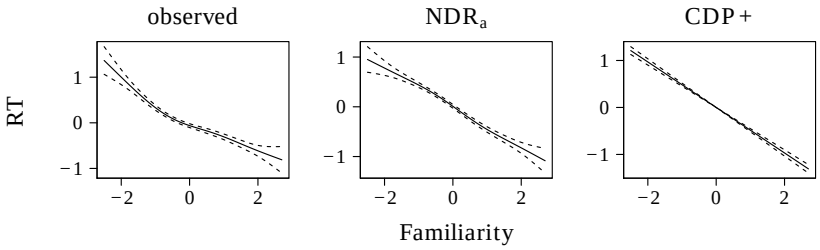

Figure 2.19. The effect of familiarity in word naming.

convex effect, however, the NDR$_a$ and CDP+ models predicts linear or near-linear effects.

2.5.5.3 Bigram frequency. Both models accurately capture the frequency effect at the word level. Frequency effects, however, also exists at a finer grain size. Baayen et al. (2006), for instance, showed an effect of bigram frequency on word naming latencies. In the NDR$_a$, bigrams have explicit representations both at the orthographic level, and many of the demi-syllable representations at the phonological level are diphones. In the CDP+, no explicit bigram representations are present. We therefore hypothesized that there might be an advantage for the NDR$_a$ over the CDP+ model for these effects.

Here, we explore the effect of two measures of orthographic bigram frequency: summed bigram frequency and mean bigram frequency. Both of these measures were predictive for the ELP naming latencies (summed bigram frequency: $t = 7.434$, $\beta = 0.146$; mean bigram frequency: $t = 11.309$, $\beta = 0.229$). The NDR$_a$ simulated the linear effect of both summed ($t = 15.675$, $\beta = 0.298$) and mean bigram frequency ($t = 26.341$, $\beta = 0.469$). Consistent with the effect of word frequency, the effect sizes in the NDR$_a$ are larger than those in the observed data. As we will clarify in the section on the overall fit of the model, however, the effects of the bigram frequency measures in the NDR$_a$ have the correct relative magnitude as compared to other lexical predictors. The CDP+ also captures the effect of mean bigram frequency ($t = 9.765$, $\beta = 0.190$) and summed bigram

frequency ($t = 3.093$, $\beta = 0.061$), although it underestimates the effect size of the summed bigram frequency effect.

Figure 2.20 shows the results of a non-linear model for mean (top row) and summed (bottom row) bigram frequency. In the observed naming latencies, there is a facilitatory effect of mean bigram frequency that increases in size for larger values of bigram frequency. The NDR$_a$ correctly captures this pattern of results, but somewhat underestimates the non-linear component of the effect. The CDP+ correctly simulates the overall facilitatory trend of the effect, but shows great uncertainty about the nature of the effect for extreme values of mean bigram frequency. The effect of summed bigram frequency is linear in the observed naming latencies. Both models correctly predict facilitation for the lowest values of summed bigram frequency, but the CDP+ and, to a lesser extent, the NDR$_a$ model incorrectly predict that the effect would level off for higher summed bigram frequencies.

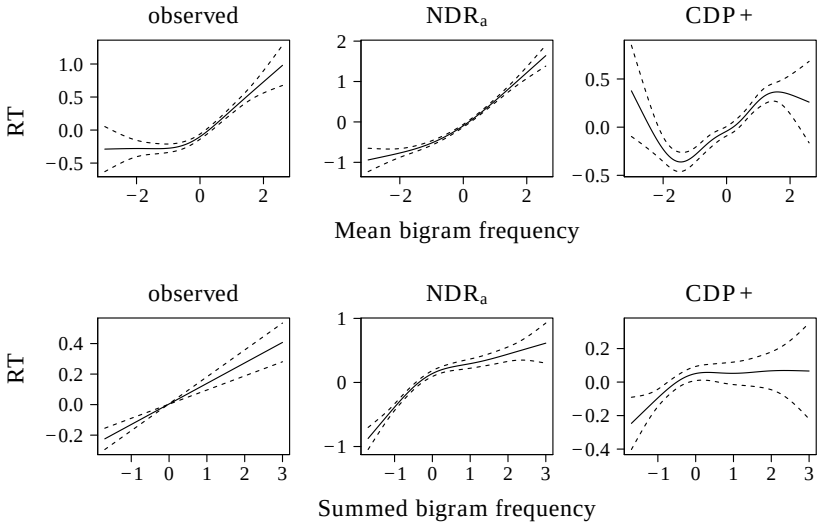

Figure 2.20. The effects of mean bigram frequency (top row) and summed bigram frequency (bottom row) in word naming.

In addition to the effect of orthographic bigram frequency, we also investigated the effect of phonological bigram frequency. The observed naming latencies showed a facilitatory linear effect of the frequency of the initial diphone (t = -6.733, β = -0.139). The NDR$_a$ (t = -4.201, β = -0.087) and the CDP+ (t = -6.641, β = -0.130) both capture this linear effect, although the NDR$_a$ underestimates the size of the effect. As can be seen in Figure 2.21, the non-linear effect is u-shaped in nature, with greater naming latencies for words with low-frequency initial diphones and - to a lesser extent - words with high frequency initial diphones. Both the NDR$_a$ and the CDP+ model successfully capture the non-linear nature of the effect. The NDR$_a$, however substantially overestimates the difficulty for words with high frequency initial diphones, whereas the CDP+ somewhat underestimates the time required to name these words. Given the sparsity of data points at the high end of the predictor range and the resulting increased width of the confidence intervals, however, strong conclusions about the performance of both models for words with high frequency initial diphones would be premature.

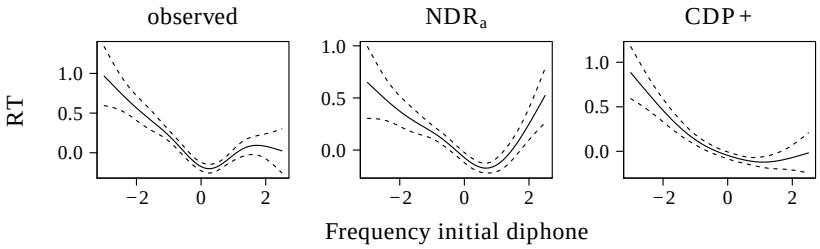

Figure 2.21. The effect of the frequency of the initial diphone in word naming.

2.5.6 Semantic effects

A final class of effects we investigated are semantic effects. First, we looked at the effect of the number of synonym sets that a word appeared in (as listed in WordNet G. A. Miller, 1990). The more different meanings a word has, the more synsets it appears in and the faster it is named (Baayen et al., 2006). Following Baayen et al. (2006), we consider two

related measures: the number of simplex synsets and the number of complex synsets. The number of simple synsets simply refers to the number of synsets a word occurs in. The number of complex synsets is defined as the number of synsets in which a word is part of a compound or phrasal unit.

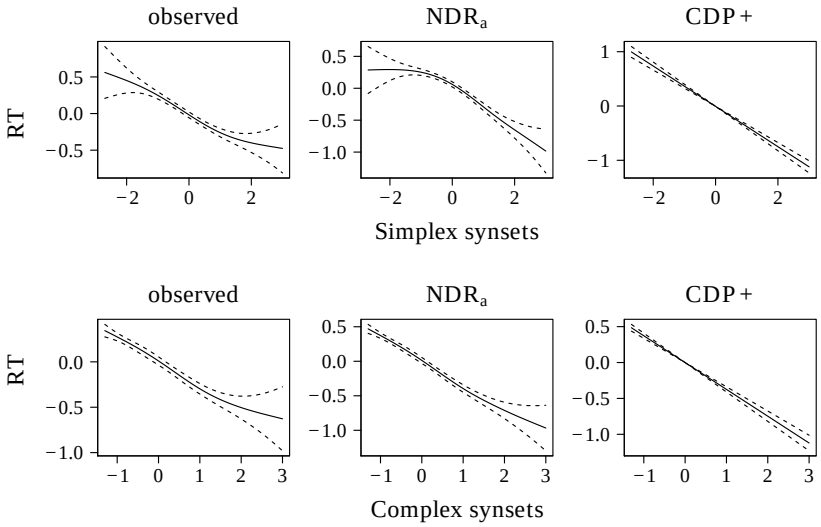

Figure 2.22. The effect of number of simplex (top row) and complex (bottom row) synsets in word naming.

Both measures have an inhibitory effect on the observed naming latencies, which is slightly larger for the number of complex synsets (t = -13.368, β = -0.268) than for the number of simplex synsets (t = -11.277, β = -0.229). The NDR$_a$ (number of simplex synsets: t = -13.541, β = -0.268; number of complex synsets: -19.407, β = -0.368) correctly simulates this pattern of results, although it overestimates the difference in effect sizes between both effects. In the CDP+ model, on the contrary, both effects are nearly identical in size (number of simplex synsets: t = -20.359, β = -0.368; number of complex synsets: -20.654, β = -0.372). Figure 2.22 presents the effect of both predictors, which are linear or near-linear in the ELP naming latencies, as well as in the simulated naming latencies

in both models. The NDR$_a$ model shows some non-linearity for words that appear in few simplex synsets, but, again, given the sparsity of data points at the lower end of the predictor range and the resulting wide confidence interval, this predicted non-linearity is not statistically robust.

A third semantic variable we looked at is morphological family size. Morphological family size is defined as the number of morphologically complex words in which a word occurs as a constituent (see, e.g., Schreuder & Baayen, 1997). Words that occur in many complex words (such as *work*) are named faster than words that occur in fewer complex words (Baayen et al., 2006). This facilitatory effect of family size was confirmed in the ELP naming latencies (t = -15.468, β = -0.306). Both the NDR$_a$ (t = -19.999, β = -0.378) and the CDP+ (t = -24.372, β = -0.425) correctly simulated this effect of family size. As can be seen in Figure 2.23, the effect of family size in a non-linear GAM is similar in the observed naming latencies and the simulations with the NDR$_a$ and CDP+ models.

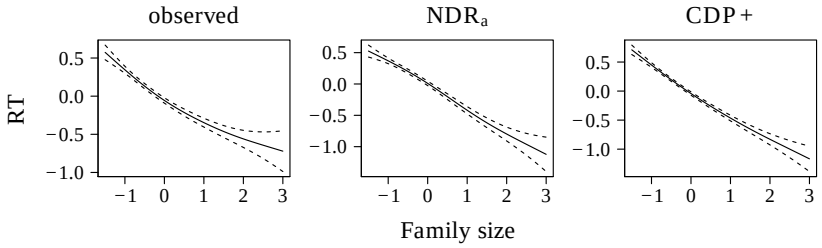

Figure 2.23. The effect of family size in word naming.

A final semantic measure is derivational entropy (Moscoso del Prado Martín, 2003). Derivational entropy is the entropy (Shannon, 1948) over the probabilities of a word's morphological family members. As such, it provides an alternative to the family size measure, with family members weighted for their token frequency. Similar to the effect of family size, derivational entropy showed a facilitatory effect in the observed naming latencies (t = -8.876, β = -0.182) that was correctly simulated in both the NDR$_a$ (t = -9.603, β = -0.194) and the CDP+ (t = -9.437, β = -0.183). In contrast to family size, however, a non-linear model of derivational

entropy on the observed naming latencies showed a fairly complex non-linear pattern. As can be seen in Figure 2.24, however, both the NDR$_a$ and the CDP+ models capture this non-linear pattern with remarkable accuracy.

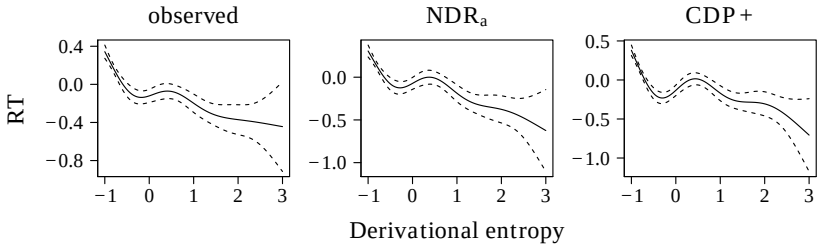

Figure 2.24. The effect of derivational entropy in word naming.

2.5.7 Overall model fit

Now that we discussed the effects of individual predictors it is time to consider the overall fit of the NDR$_a$ model. A first issue to address is the item-level performance (see, e.g., Spieler & Balota, 1997) of the models. For the current set of 2,524 monosyllabic nouns, the correlations between the observed naming latencies and the naming latencies simulated by the NDR$_a$ ($r = 0.500$) and CDP+ ($r = 0.492$) were similar. The CDP+ uses 25 parameters to obtain this performance. By contrast, the NDR$_a$ has only 8 free parameters. Therefore, the AIC (Akaike, 1974) score of the NDR$_a$ (6453.48) is much lower than that of the CDP+ model (6515.81). These AIC scores indicate that the NDR$_a$ model is 93,018,468,769,241 times more probable than the CDP+ model (Akaike, 1980).

2.5.7.1 Predictor effect sizes. A second issue that concerns the overall fit of the model are the relative effect sizes of the different predictors. Figure 2.25 plots the modeled predictor coefficients (βs) in the linear regression models for each predictor in the observed data against the coefficients in the naming latencies simulated by the NDR$_a$ (top panel) and CDP+ (bottom panel) models. Ideally, the points in these graphs

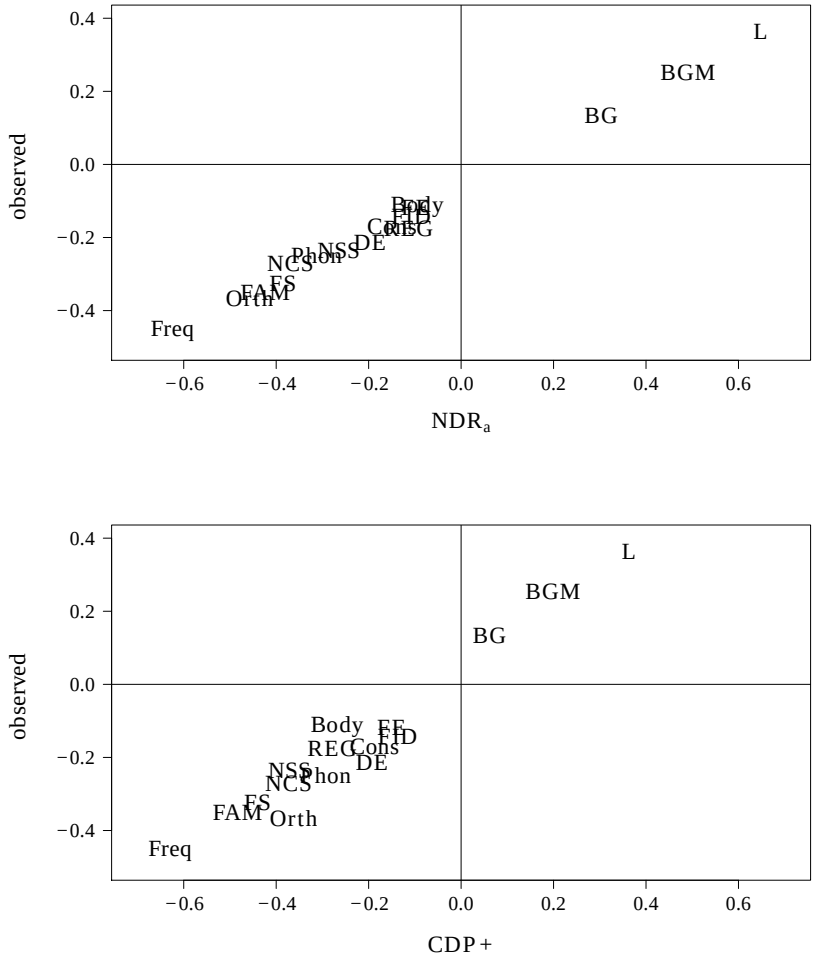

Figure 2.25. Comparison of predictor coefficients for the observed data and the simulations of the NDRa (top panel) and CDP+ (bottom panel) models. Predictors from bottom to top: Freq (frequency), Orth (orthographic neighborhood density), FAM (familiarity), FS (family size), NCS (number of complex synsets), Phon (phonological neighborhood density), NSS (number of simplex synsets), DE (derivational entropy), REG (regularity), Cons (consistency), FID (frequency initial diphone), FE (friends-enemies measure), Body (body neighborhood density), BG (summed bigram frequency), BGM (mean bigram frequency), L (length).

are on a straight line. This would indicate that the relative effect sizes in the simulated data are identical to those in the observed data. In the plot for the NDR$_a$, simulated coefficients deviate very little from this ideal pattern of results. The accuracy of the predictor effect size in the NDR$_a$ is confirmed by a correlation of $r = 0.997$ between the coefficients for the observed data and the coefficients in the NDR$_a$ simulations. The CDP+ simulated coefficients also show a high correlation with the coefficients for the observed data ($r = 0.972$). Nonetheless, the relative effect sizes deviate more from those in the observed data for the CDP+ model than for the NDR$_a$.

The CDP+ model coefficients show two particular problems with the relative effect sizes in the CDP+ simulations. First, the effect sizes of the neighborhood density measures are too similar in the CDP+. There is an overestimation of the effect of body neighborhood density and an underestimation of the effect of orthographic neighborhood density. Second, the effect of regularity is substantially larger than that of consistency. This stands in contrast to the observed data, where both effects are very similar in size. These observations indicate that the CDP+ model puts too much importance on processes underlying the effects of body neighborhood density and regularity.

The effect sizes for the NDR$_a$ and, to a lesser extent, the CDP+, are larger than those for the observed data. Importantly, this does not imply that the models are overfitting predictor effects. As noted by Adelman and Brown (2008), the standard deviation for modeled naming latencies is smaller than that for observed latencies. The reason for this is that models operate under perfect noise-free conditions. This stands in sharp contrast to the observed naming latencies, even when those observed latencies are averaged over participants. We normalized the observed and simulated latencies prior to our simulations. As a consequence, the smaller standard deviation in the simulated data results in larger estimated effect sizes. The increased effect sizes in the NDR$_a$ as compared to the observed data, therefore, are a result of the noise-free conditions in the model simulations.

2.5.7.2 Principal components regression analysis. A third issue regarding the overall model fit is how well the model characterizes the multidimensional structure described by the predictors under investigation. In the predictor simulations we fitted separate models for each predictor. Although this allowed us to get away from the multicollinearity problem, it implies that the effect for any given predictor may be confounded with that of another predictor. We therefore sought to verify that the overall characterization of the multidimensional predictor space by the NDR$_a$ is correct.

Table 2.3 shows the results of a linear regression model fit on the first eight principal components of the 16-dimensional space described by our predictors. Together, these eight principal components explained 86% of the variance in the input space. The NDR$_a$ predicts the right sign for 7 of the 8 principal components. The only component the NDR$_a$ has problems with is PC4, which has high positive loadings for neighborhood density measures as well as bigram frequency measures. For this principal component the NDR$_a$ incorrectly predicts an inhibitory effect, while the observed data show no such effect. Consistent with the effect sizes for the predictors themselves, the effect sizes of the principal components are larger for the NDR$_a$ simulated latencies than for the observed data. Again, however, the relative magnitude of the effect sizes (βs) is highly similar for the simulated and observed data ($r = 0.94$). This demonstrates that the NDR$_a$ simulations capture the overall input space quite well.

The CDP+ model does not capture the input space as well as the NDR$_a$. This is reflected in a somewhat lower correlation with observed principal components coefficients ($r = 0.86$). The CDP+ model correctly predicts that there should be no effect of PC4, but incorrectly predicts an inhibitory effect for PC3, which has strong negative loadings for consistency and friends minus enemies. Furthermore, the CDP+ model fails to capture the inhibitory effect of PC7, which contrasts regularity (high positive loading) with consistency and the frequency of the initial diphone (high negative loadings).

Table 2.3. Results of a principal components analysis on the 16 dimensional space described by the predictors. Listed are t-values and β coefficients for the first 8 principal components.

	observed		NDR_a		CDP+	
	t	β	t	β	t	β
PC1	-26.827	-0.235	-52.651	-0.312	-43.922	-0.304
PC2	-6.985	-0.070	-20.345	-0.139	-3.659	-0.029
PC3	-1.745	-0.020	-15.654	-0.123	3.435	0.031
PC4	-0.059	-0.001	8.058	0.078	0.014	0.000
PC5	2.864	0.050	10.858	0.128	11.677	0.160
PC6	7.396	0.141	9.780	0.126	8.459	0.127
PC7	7.052	0.139	19.376	0.259	1.079	0.017
PC8	3.329	0.078	10.169	0.161	5.302	0.098

2.5.8 Comparison to a dual-route architecture

The single route architecture of the NDR_a model provides a good fit to observed reading aloud data. It could be the case, however, that adding a non-lexical route would improve the model's performance. This issue is particularly relevant given the fact that the non-lexical route of the CDP+ model has a significant contribution in terms of explained variance, both in word and non-word naming (Perry et al., 2007). To resolve this issue we implemented a sub-lexical route by means of a Rescorla-Wagner network that learned to associate orthographic input cues (letters and letter bigrams) with phonological outcomes (demi-syllables). We trained this non-lexical Rescorla-Wagner network on the same set of training data as the NDR_a. This resulted in three additional model components, describing the activation of the first ($ActPhonSub_1$) and second ($ActPhonSub_2$) demi-syllable through the non-lexical route and the entropy over these activations ($HSub$). We then fitted two linear regression models to the (inverse transformed) observed naming latencies. The first linear model included as predictors the (log-transformed) components of the original NDR_a model. The coefficients of this linear model were similar to the parameter settings used in the simulations throughout this chapter

Table 2.4. Results of a linear model predicting observed reaction times from model components. Listed values are component t-values.

	NDR_a	NDR_a^2
Lexical route		
ActLexeme	5.011	3.231
ActPhon_1	5.989	6.003
ActPhon_2	12.259	11.499
H	7.520	7.077
Complexity	18.019	16.851
Non-lexical route		
ActPhonSub_1	NA	0.398
ActPhonSub_1	NA	1.114
HSub	NA	1.246

($r = 0.987$). The second linear model included as predictors not only the components of the NDR_a model, but also the 3 additional measures derived from the sub-lexical route.

Table 2.4 presents the t-values associated with each component in the linear model containing the lexical components of the NDR_a and the linear model containing both lexical and sub-lexical components. This dual-route model will henceforth be referred to as the NDR_a^2. Table 2.4 shows that the relative contributions of the lexical components are similar in the NDR_a and NDR_a^2. Adding a sub-lexical route to the model architecture, therefore, does not affect the contribution of the lexical model components much. In addition, the sub-lexical components in the NDR_a^2 do not improve the explanatory power of the model. Neither the activation of the demi-syllables from the orthography, nor the entropy over these activations reaches significance in the linear model for the NDR_a^2. Furthermore, the predicted values of the NDR_a and NDR_a^2 linear models are highly similar ($r = 0.997$), and both models show similar correlations with the observed naming latencies (NDR_a: $r = 0.483$; NDR_a^2: $r = 0.484$).

The results for the linear models presented here demonstrate that the addition of a non-lexical route does not improve the performance of the NDR_a in word naming. In addition, the simulations for the individual predictors demonstrated that the effects documented in the non-word naming literature are adequately captured by the single lexical route architecture of the NDR_a. The current simulations therefore suggest that a single-route architecture is sufficient to capture the patterns of results observed in both word and non-word naming experiments.

2.5.9 Non-word frequency effect

A reanalysis of the McCann and Besner (1987) naming latencies for non-words provides independent evidence for the use of a lexical architecture in non-word naming. For each of the 154 non-words in the study, we obtained unigram frequencies from the Google 1T n-gram corpus (Brants & Franz, 2006). The database of google unigram frequencies only includes words with a frequency of 200 or greater. It is striking therefore, that only 14 of the 154 non-words did not appear in the Google unigram corpus. A Google web search for these 14 words showed that even the least frequent of these words still appeared on 7,700 web pages. Furthermore, the average google unigram frequency of non-words (197,396) is comparable to that of low frequency English words like *scrum* (frequency: 196,879) or *minstrel* (frequency: 196,617). This suggests that the distinction between words and non-words is not as absolute as is commonly believed. As for real words, any given non-word therefore may or may not have a representation in the mental lexicon of an individual language user. The probability of such a representation existing is a function of the frequency of the word or non-word.

Given these observations we investigated whether there was a frequency effect of non-words in the naming latencies for experiment 1 in McCann and Besner (1987). We found a highly significant effect of non-word frequency (t = -6.054, β = -0.441). This effect of non-word frequency existed over and above the effects of word length, orthographic neighborhood density, base word frequency and non-word type (regular or pseudo-homophone). Non-word frequency was the most powerful pre-

dictor of non-word naming latencies and showed a correlation to observed naming latencies ($r = 0.44$) similar to that of the word frequency measure in the ELP naming latencies for real words ($r = 0.45$).

To verify that the architecture of the NDR_a supports non-word frequency effects, we retrained the model on an input set that contained the non-words nouns from the McCann and Besner (1987) study with appropriate relative frequencies. We embedded these non-words in word trigrams through frequency weighted sampling from trigrams that contained nouns and replacing the nouns in these trigrams with the non-word nouns. The resulting model is a truly lexical model of non-word naming, in which non-words are read in exactly the same way as real words. With parameter settings identical to those in all previously reported simulations, this model correctly simulated the non-word frequency effect (t = -7.880, β = -0.539). As expected, the CDP+ model does not capture this effect (t = -1.744, β = -0.140), although we do see a non-significant trend in the expected direction, presumably due to the strong correlation between non-word length and non-word frequency ($r = -0.480$).

The results of a non-linear model for non-word frequency are presented in Figure 2.26. The observed naming latencies show a facilitatory effect that levels off for the highest frequency non-words. The NDR_a captures the facilitatory trend, but predicts the effect is somewhat concave rather than convex in nature, with an effect that levels off for the 14 non-words that did not appear in the Google unigram corpus. Given the limited size of the current set of non-words, we are hesitant to draw strong

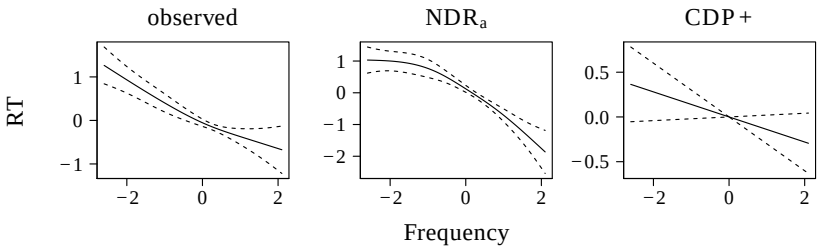

Figure 2.26. The effect of frequency in non-word naming.

conclusions regarding this discrepancy. If future research were to indicate that the observed effect is robust and that the NDR$_a$ systematically underestimates its non-linearity, we hypothesize that revised, more carefully selected training data might lead to better simulation results. The CDP+ simulations, finally, show the non-significant trend mentioned above.

The non-word frequency effect suggests that the dichotomous distinction between words and non-words is perhaps better thought of as a difference on a gradient scale, with high frequency words on one end of the scale and low frequency words on the other. Such a gradient scale fits well with the architecture of the NDR$_a$, in which the difference between word and non-word processing is quantitative rather than qualitative in nature: words and non-words are processed by the same cognitive architecture, with differences only in the amount of activation flowing through the system.

We conclude this section on a note about the quantitative performance of the NDR$_a$ and CDP+ models for the McCann and Besner (1987) naming latencies. The predicted naming latencies of both the NDR$_a$ ($r = 0.092$) and the CDP+ model ($r = 0.101$) correlate poorly with the observed naming latencies. One potential explanation for the poor quantitative performance of both models may be the fact that the stimulus list of experiment 1 of McCann and Besner (1987) consisted of non-words only. Task strategies, therefore, may substantially differ from experiments in which mixed stimulus lists are used. Alternatively, the visual input interpretation mechanism used in the current implementation of the NDR$_a$ may be too simplistic. The predicted values of a simple linear model including a frequency-weighted version of the visual complexity measure ($Complexity/(LogFrequency$ + back-off constant)) rather than the original complexity measure boosted the correlation with the observed non-word naming latencies to $r = 0.489$, a correlation similar to the correlation between the word naming latencies in the ELP and the word naming latencies simulated by the NDR$_a$. We return to the issue of familiarity with the visual input in the discussion section.

2.5.10 Pronunciation performance

The processes by which responses are learned are relatively well understood and large-scale linguistic corpora provide us with realistic input to these processes. As demonstrated in the simulations reported thus far, the discriminative learning algorithm that underlies the NDR$_a$ provides a precise and powerful explanation of these bottom-up processes and their behavioral manifestations in observed naming latencies. What the discriminative learning core of the NDR$_a$ model does not do, however, is generate actual pronunciations.

The selection of the appropriate target response is perhaps best thought of as a response conflict resolution task. In the words of Ramscar, Dye, Gustafson and Klein (2013), "Response conflict will arise whenever the requirements in a specific task conflict with an equally or more strongly learned pattern of responding that is prompted by the same context. To successfully resolve this conflict, an individual must be able to effectively override the biased response in favor of a less well-learned (or less well-primed) response that is more appropriate to the context" (see also Yeung et al., 2004; Novick et al., 2010). In the NDR$_a$ model a response conflict arises whenever a non-target demi-syllable receives a higher activation than the target demi-syllables.

Response conflicts are typically resolved by a top-down verification mechanism that integrates the activated responses with the context of the current task. Dell (1986) and Levelt et al. (1999), for instance, proposed such top-down verification mechanisms in their models of language production. In reading aloud, the task of a top-down checking mechanism is to find out which of the activated phonological units should be pronounced given the visual presentation of a word or non-word. What we suggest, therefore, is that there is a functional separation between the bottom-up linguistic support for phonological units that arises in the discrimination learning networks that form the linguistic core of the NDR$_a$ model and the top-down verification mechanism that evaluates the appropriateness of these phonological units given the task of naming the presented word or non-word.

There is a wealth of evidence in both the neuroscience and reading literatures to support a functional separation of this kind (see, e.g., Yeung et al., 2004). In particular, the anterior cingulate cortex (ACC) and the pre-frontal cortex (PFC) seem to play an important role in resolving competition between different potential responses (see, e.g., Botvinick et al., 2004). Functionally, the ACC appears to serve as a detector, monitoring conflict between candidate responses and activating areas in the PFC that facilitate the selection of the appropriate target response when conflicts arise.

The Stroop task, in which subjects have to name the text color of an orthographic representation of a conflicting color word neatly illustrates the dynamics of this process in a reading task. When the word "blue" is printed in red, the correct response is "red". In literate adults, however, the orthographic activation of "blue" interferes with the correct response. In the Stroop task, activation in the PFC, and in particular the left inferior frontal gyrus has been shown to reflect the effort required to produce the text color "red" rather than the strongly activated competitor "blue" (Milham et al., 2003). As noted by Novick et al. (2010), the PFC plays a functionally similar role when response conflict arises in a range of more straightforward lexical tasks, including lexical decision (see e.g., Grindrod et al., 2008), verb generation (Thompson-Schill et al., 1997), picture naming (Kan & Thompson-Schill, 2004), and phonological and semantic judgment tasks (Snyder et al., 2007), as well as when interpretative conflicts arise during normal reading (Novick et al., 2010).

As we noted above, the processes by which responses are learned are relatively well understood. By contrast, a lot of uncertainty remains about how exactly the top-down verification processes in the pre-frontal cortex that select the appropriate response from a set of activated potential responses work. The main objective of the present study is to demonstrate that the discrimination learning networks in the NDR_a model capture important aspects of the bottom-up learning processes and their manifestations in observed naming latencies, much like the original NDR model captures a wide range of reaction time effects in the lexical decision task. Given the increased prominence of the actual response in the

reading aloud task as compared to the lexical decision task, however, it is important to explicitly demonstrate that the architecture of the NDR$_a$ model is fully compatible with a top-down checking mechanism that generates concrete and plausible pronunciations. We therefore present an implementation of such a verification mechanism, as well as the word and non-word naming performance of the NDR$_a$ model when such a checking mechanism is added on top of the discrimination learning networks of the model.

2.5.10.1 Response conflict resolution in the NDRa. Given our limited understanding of the functional architecture of the PFC, a considerable amount of uncertainty remains with respect to the optimal implementation of a verification mechanism. The checking mechanism proposed here is a crude first approximation of what we think the architecture of a checking mechanism might look like. The basic rationale behind this checking mechanism is that the PFC filters the set of lexical representations that activate demi-syllables to only include the subset of lexemes that share orthographic features with the target word or non-word. As such, the checking mechanism used here limits response conflict monitoring to the lexical level: the pre-frontal cortex monitors the set of activated lexemes and removes from this set those lexical representations that are inappropriate given the context of the orthographic input. No further response monitoring takes place at the phonological level.

Our implementation of the checking mechanism builds on the idea that a lexeme points to letters and letter order information (which is required, for instance, for writing). This information is then compared on-line against the orthographic features in the input. As pointed out earlier, the assumption that language users are able to compare the orthographic features associated with a lexeme to the orthographic features in the input is not unique to the verification mechanism proposed here. Instead, it is a general assumption of discrimination learning that is necessary to evaluate whether or not the outcome of a learning event is predicted correctly, and that is consistent with theories of cortical processing that

propose a bi-directional pass of information between higher and lower levels of information (Friston, 2005).

How exactly does the checking mechanism work? Consider the example word *bear*. When the orthographic string *bear* is presented on the screen, activation spreads to a large number of lexical representations. The set of activated lexemes includes orthographic neighbors of *BEAR* such as *PEAR*, *HEAR* and *FEAR* as well as the target lexeme *BEAR* itself. For a correct pronunciation of the word *bear*, however, it is sufficient to consider only those demi-syllables that are activated by the target lexeme *BEAR*. The checking mechanism therefore limits activation of demi-syllables to those units that are activated by this target word lexeme. In the case of *bear*, the initial demi-syllable that receives most activation from the lexeme *BEAR* is *b8*, whereas the most active second demi-syllable is *8R*. The model therefore correctly pronounces the word *bear* as *b8R*.

For a vast majority of all words, the most active word-initial and word-final demi-syllables are compatible in the sense that the vowel in the initial and final demi-syllables is identical. For 8 out of the 2,524 monosyllabic words in our data set, however, the vowel in the most active first and second demi-syllable are different. In these cases the checking mechanism gives preference to the vowel in the second demi-syllable. This implementational decision corresponds to the fact that the activation of the second demi-syllable has a somewhat higher weight in the NDR_a as compared to the activation of the first demi-syllable (weight $ActPhon_1$: 0.43, weight $ActPhon_2$: 0.49) and to the increased perceptual prominence of rhymes as compared to onset plus vowel sequences.

For a non-word such as *braint* no lexical representation exists. Limiting the phonological units that influence pronunciation to those activated by the target word lexeme therefore does not work for non-words. Instead, the checking mechanism needs to identify which lexical representations share relevant orthographic features with the non-word *braint*. Only the phonological activation generated by these lexemes should influence non-word pronunciation. The question then becomes how to define the term "relevant orthographic features". One option is to include all lexemes

whose orthographic representations share at least n orthographic bigrams with the non-word presented on the screen. The problem with such a definition is that the checking mechanism would be relatively insensitive to the serial nature of the non-word naming task.

We propose an alternative definition that takes into account the left-to-right nature of the non-word naming task by varying the set of lexemes that influence pronunciation in a serial manner. For the first demi-syllable, the checking mechanism proposed here ensures that the initial demi-syllables considered for pronunciation are restricted to the set of initial demi-syllables that receive activation from lexemes that share the orthographic onset with the presented non-word. For the non-word *braint*, for instance, only those initial demi-syllables that are activated by one of the 69 lexemes that share the orthographic onset *br* with the lexeme *BRAINT* are considered for pronunciation (e.g., *BRACE*, *BRONZE*, *BRUSH*, ...).

As for existing words, the checking mechanism gives preference to the vowel in the second demi-syllable over the vowel in the first demi-syllable. To allow for a correct pronunciation of non-words the checking mechanism therefore considers the combination of the orthographic vowel and coda when selecting the appropriate second demi-syllable (i.e., it limits the set of word-final demi-syllables considered for pronunciation to the set of word-final demi-syllables that receives activation from lexemes that share the orthographic rhyme with the presented non-word).[6] For the non-word *braint*, for instance, only those word-final demi-syllables that are activated by the lexemes *FAINT*, *PAINT*, *PLAINT*, *SAINT* and *TAINT* are considered for pronunciation.

Consistent with the architecture of the NDR_a we weighted the contribution of lexical representations to demi-syllable activations for the amount of activation they received from the visual presentation of the non-word (see Equation 2.3). For the non-word *braint* the initial demi-syllable that received the highest activation from the co-activated lexical repres-

[6] This assumes that the system is sensitive to the distinction between vowels and consonants. A similar assumption is made in the non-lexical route of the CDP+ model, which parses the visual input into consonant and vowel slots in a grapheme buffer.

entations was *br*1, whereas the highest activated second demi-syllable was 1*nt*. Together, these demi-syllables yield the correct pronunciation of the non-word *braint*, which is *br*1*nt*. The same procedure was used to resolve ties for existing words for which the activation of a demi-syllable from the target word lexeme was equally high for two or more demi-syllables (i.e., for 64 word-initial demi-syllables (2.54%) and 184 word-final demi-syllables (7.29%)).

For a vast majority of the 2,524 words and 1,822 non-words under consideration, the algorithm described above yields a single most highly activated first and second demi-syllable. For 65 words (2.58%) and 68 non-words (3.73%), however, two or more potential word-final demi-syllables still receive equal activation. For these non-words the checking mechanism resorts to the phonological activations generated by the set of lexical representations that share only the orthographic coda with the non-word to resolve the tie, considering only those word-final demi-syllables that share the phonological coda with one of the demi-syllables involved in the tie. Ties are much less common for word-initial demi-syllables than for word-final demi-syllables: for 0 words and 2 (0.11%) non-words two or more potential word-initial demi-syllables receive equal activation. Given that initial demi-syllables were already selected on the basis of orthographic overlap with the onset only, we resolved these two ties by comparing the overall activation of initial demi-syllables that shared the phonological onset with one of the demi-syllables involved in the tie in the system as a whole (i.e., for all lexemes in the lexicon) given the orthographic features of the non-word. This backup procedure eliminates all ties and ensures that the model yields unique pronunciations for all words and non-words.

For the naming latency simulations reported thus far, we used a training corpus that consisted of a little under 9 million two and three word phrases from the British National Corpus (BNC). While this corpus proved sufficiently large to obtain a good quantitative fit to the observed naming latencies it covered only 3,209 of all 3,908 uninflected mono-morphemic words in our phonological training lexicon (82.11%). A complete coverage of the mono-morphemic word space, however, is essential for optimal

non-word pronunciation performance in the NDR_a: the more orthographic neighbors of a non-word are in the training data, the higher the chance that the model will arrive at the correct pronunciation of that non-word. To improve the coverage of the training data we therefore retrained the NDR_a on all trigrams from a $1,000,000,000$ word Usenet corpus that consisted of words from a precompiled list of $240,534$ English words and proper nouns, using orthographic bigrams as input cues. This resulted in a training set that comprised $134,684$ word types, which covered $3,901$ of the $3,908$ uninflected mono-morphemic words in the CELEX lexical database (99.82%).

Furthermore, the training data for the lexeme-to-phonology network used in the simulations were restricted to lowercase mono-morphemic, mono-syllabic words that were present in the CELEX lexical database, that consisted of at least 3 letters and for which word frequencies were available in the ELP. These restrictions led to the exclusion of the mono-syllabic words that are crucial for the correct pronunciation of non-words beginning with *ps-* (*s*) or ending with *-ach* ({*k*), *-if* (*If*), *-ewn* (*5n*), *-eich* (*1J*) or *-udd* (*Vd*). We therefore added the words *psalm* (only plural present in CELEX), *if* (less than 3 letters), *mach*, *reich* and *ludd* (capitalized in CELEX), and the past tense forms *hewn*, *sewn*, *strewn* and *shewn* (not mono-morphemic) to the training data for the lexeme-to-phonology network with their respective frequencies in the ELP. No other mono-morphemic, mono-syllabic words beginning with *ps-* or ending with *-ach*, *-if*, *-ewn*, *-eich* or *-udd* that were not already in the training data were present in the CELEX lexical database. Finally, whereas in the naming latency simulations reported above we limited the set of lexical neighbors to words that can be used as nouns for computational efficiency, we allowed all words for which lexemes were present in the lexeme-to-phonology network to enter the equation when generating simulated pronunciations. As such, the current training data provided the model with the optimal conditions for correct pronunciation performance.

2.5.10.2 Simulation results. The NDR$_a$ model generated correct pronunciations for 2,504 of the 2,524 monosyllabic words in our database, resulting in a word pronunciation performance of 99.21%. A majority of the pronunciation errors (14 out of 20) concerned words that have more than one pronunciation in the CELEX lexical database, such as *tear* or *wind*. For these words the model chooses the more frequent pronunciations *t8R* and *wInd* over the less frequent pronunciations *t7R* and *w2nd*, as would participants in a reading aloud task. Of the remaining 6 erroneous pronunciations in the NDR$_a$, 5 were based on position-specific grapheme-to-phoneme conversions that exist in other English words: *blouse* is pronounced as *bl6s* rather than *bl6z* (analogous to *house* (*h6s*)), *smith* as *smID* rather than *smIT* (analogous to *with* (*wID*)), *draught* as *dr#t* rather than *dr#ft* (analogous to *fraught* (*fr$t*)), and *queer* and *weir* are pronounced without the final *R*: *kw7* and *w7* (analogous to *their* (*D8*)). The remaining erroneous pronunciation contains a grapheme-to-phoneme conversion that is not attested in English: the NDR$_a$ pronounces *zone* as *z5ks*.

The pronunciations of the CDP+ model differed from those in the CELEX lexical database for 300 out of the 2,524 words in our database of monosyllabic words. Importantly, however, a substantial number of these differences may be due to differences in the training data. While the sub-lexical route of the CDP+ model was trained on the British pronunciations in the CELEX lexical database, the training data for the interactive activation model in the lexical route are not explicitly specified in Perry et al. (2007). The pronunciations of the CDP+ model, however, suggest that the lexical route was trained on a variety of American English, rather than British English.

Nearly half the differences (140 out of 300) between the pronunciation of the CDP+ model and the phonological representation of these words in CELEX, for instance, concern the use of the vowel { rather than the vowel # (e.g., *dance* pronounced as *d{ns* rather than *d#ns* or the use of the vowel *9* rather than the vowel *$* (e.g., *pork* pronounced as *p9k* rather than *p$k*). Similarly, a large number of differences (103) concerns the omission of a word-final *R* in the pronunciation (e.g., *beer* is pronounced

as *b7* rather than *b7R*). Furthermore, 13 "mispronunciations" concern the use of *Z* rather than _ at the end of words (e.g., *range* is pronounced as *r1nZ* rather than *r1n_*). Given the fact that these pronunciations are likely to reflect differences in training data rather than differences in model performance, we decided to not consider these pronunciations erroneous.

After discounting the differences in pronunciation that may be due to differences in the training data, 44 pronunciation errors remain for the CDP+ model. The pronunciation performance of the model therefore is 98.26%. As for the NDR_a model, a substantial part of these errors concerned words that have more than one pronunciation in CELEX (14 out of 44 errors). For 18 out of the 30 remaining mispronunciations, pronunciations generated by the CDP+ model used position-specific existing grapheme-to-phoneme conversions (e.g; *cyst* was pronounced as *kIst* rather than *sIst*, *dough* was pronounced as *d5f* rather than *d5*). The final 12 errors contained grapheme-to-phoneme conversions that are not attested in English (e.g., *steppe* was pronounced as *stEpt* rather than *stEp*, *ewe* was pronounced as *jju* rather than *ju*).

In summary, both models yield accurate pronunciations for words. After discounting for potential differences in training data and words with more than one pronunciation, the NDR_a model mispronounced only 6 words while the CDP+ model made 30 pronunciation errors. Word pronunciations, however, are generated by lexical architectures in both the NDR_a and CDP+ model. By contrast, for non-word pronunciations the CDP+ uses a learning network that directly maps orthographic units onto phonological units, whereas the NDR_a relies on co-activation of orthographic neighbors in a lexical architecture. Much more than word pronunciation, therefore, non-word pronunciation provides a litmus test for the single-route lexical architecture of the NDR_a model.

The NDR_a model generated pronunciations identical to those in the ARC non-word database (Rastle et al., 2002) for $1,289$ of all $1,822$ non-words in our non-word data set (70.75%). The CDP+ model performed similarly and yielded pronunciations identical to those in the ARC database for $1,278$ of all $1,822$ non-words in our data set (70.14%).

As for word pronunciations, a simple comparison of the pronunciations of both models with the pronunciations in the ARC non-word database, however, may not be the best way to evaluate non-word pronunciation performance. Differences exist between the phonological encoding in the ARC non-word database and the training data of both models. This results in pseudo-erroneous responses, in which the model generates a pronunciation that differs from the pronunciation in the ARC non-word database, but that is correct given the model's learning experience. In addition, multiple possible pronunciations exist for a substantial percentage of non-words. The non-word *zeaf*, for instance, can be pronounced as *zif* (analogous to *leaf*, CELEX frequency: 270) or as *zEf* (analogous to *deaf*, CELEX frequency 183). This ambiguity is reflected in the considerable between-subject variation in non-word pronunciation experiments. Following Perry et al. (2007) we therefore adopted a lenient error scoring criterion similar to that proposed by Seidenberg et al. (1994), according to which a non-word pronunciation is correct if it is based on grapheme-to-phoneme conversions that exist in real English words. The lenient scoring criterion used here, however, is a bit stricter than that proposed by Seidenberg et al. (1994), in the sense that we considered non-word pronunciations as correct if and only if the orthography-to-phonology mapping for the onset, vowel and coda existed for a monosyllabic word in CELEX.

Using the lenient scoring criterion, the NDR_a mispronounced only 38 non-words, for a non-word pronunciation performance of 97.91%. Out of these 38 erroneous responses, 5 concerned mispronunciations of the onset (e.g., *lafe* pronounced as *T1f* rather than *l1f*), 6 concerned mispronunciations of the vowel (e.g., *beik* pronounced as *bUk* rather than *b1k*) and 27 concerned mispronunciations of the coda (e.g., *brelte* pronounced as *brEt* rather than *brElt*).

By contrast, 250 pronunciations of the CDP+ model are classified as erroneous when the lenient scoring criterion is used. The non-word pronunciation performance of the CDP+ model, therefore, is 86.28%. Consistent with the observations of Perry et al. (2007), a large percentage of the pronunciation errors displayed the pattern that a phoneme was

missing from the pronunciation. Perry et al. (2007) state that reducing the naming activation criterion (i.e., the threshold activation for pronouncing a phoneme) from 0.67 to 0.50 substantially reduces the number of erroneous pronunciations in the model. Indeed, changing the naming activation criterion parameter from 0.67 to 0.50 results in correct pronunciations for 64 out of the 250 mispronounced non-words (25.60%). Nonetheless, even with this change in the naming activation criterion parameter, 55 words are pronounced incorrectly due to missing phonemes in the output.

Of the remaining 131 errors in the pronunciations of the CDP+ model, 58 concerned the pronunciation of a word-initial *ps−* as *p*. This seems to suggest that similar to the original training data for the NDR_a, the training data for the CDP+ model did not contain the orthography-to-phonology conversion from *ps-* to *s*. In addition, the CDP+ model incorrectly inserted a word-final *-d* in 13 cases, most likely due to competition from paste tense forms (e.g., *rhuin* pronounced as *rund* rather *run*). The NDR_a was not susceptible to these types of errors, because no past tense forms were present in the training data for the lexeme-to-phonology network. Furthermore, as for real words, word-final *-nge* was pronounced as *Z* rather than _ for 14 non-words. After adjusting the naming activation criterion parameter downwards and excluding errors that were likely due to differences in the training data, 46 erroneous responses remain for the CDP+ model, for an adjusted non-word naming performance of 97.48%. Most commonly, word-final *-nc* clusters were mispronounced (20 cases, e.g., *rhanc* pronounced as *r{nk* rather than *r{Nk*).

Overall, the NDR_a model performs at least as well at non-word pronunciation as does the CDP+. Both the number of pronunciations that are identical to those in the ARC non-word database and the number of correct pronunciations when a more lenient scoring criterion is used are slightly higher in the NDR_a model than in the CDP+ model. This demonstrates that the single-route architecture of the NDR_a allows for highly competitive non-word pronunciation performance.

2.6 Discussion

2.6.1 Single-route architecture

The use of a single, rather than a dual-route architecture is perhaps the most important aspect of the current work. The DRC (M. Coltheart et al., 2001), CDP (Zorzi et al., 1998b; Perry et al., 2007, 2010) and triangle model (Seidenberg & McClelland, 1989; Plaut et al., 1996; Harm & Seidenberg, 2004) all are dual-route models of reading aloud. Here, we presented a new single-route model of reading aloud that is based on the equilibrium equations (Danks, 2003) for the learning algorithm of Rescorla and Wagner (1972). We demonstrated that this single route model replicates a wide range of effects that have been documented in the experimental literature and shows an overall fit to the data comparable to or better than that of the most successful dual-route model. Furthermore, we showed that adding a sub-lexical route to the model did not improve its performance.

While the single versus dual route debate remains as open in the neuroscience literature as it is in the functional level linguistics and cognitive science literature, the single-route architecture of the NDR_a is consistent with the results of a large number of studies in the neuroscience literature. These studies found activation of the same brain regions in word and non-word reading, with no unique brain regions that are active in non-word reading only (see Wydell et al., 2003; Wilson et al., 2005; Church et al., 2011; Rumsey et al., 1997; cf. Jobard et al., 2003 for examples of conflicting evidence). These findings do not fit very well with dual-route models, in which qualitatively different processes underlie word and non-word reading. Rather than physically different architectures for word and non-word naming, differences in the timing (Cornelissen et al., 2003; Wilson et al., 2005; Juphard et al., 2011) and intensity (Wilson et al., 2005) of the activation of the *same* brain regions were observed between word and non-word reading. As noted by Wilson et al. (2005, p.1), for instance, "relative to words, pseudo-words elicit more robust activation in the left inferior temporal gyrus (ITG, see e.g., Price et al.,

1996; Brunswick et al., 1999; Paulescu et al., 2000; Xu et al., 2001) and the left inferior frontal gyrus (IFG, see e.g., Herbster et al., 1997; Rumsey et al., 1997; Brunswick et al., 1999; Fiez et al., 1999; Hagoort et al., 1999; Paulescu et al., 2000; Xu et al., 2001; Binder et al., 2003)".

There are, however, some outstanding issues that warrant further discussion. First, the NDR$_a$ assumes that processing is strictly parallel, while a number of experimental findings suggest that at least some serial processing occurs when preparing to read aloud words and non-words. Second, we made decisions regarding the grain size of representations at both the orthographic (letters and letter bigrams) and the phonological (demi-syllables) level that proved adequate for the current purposes but that are likely to be an oversimplification of more complex neural structures. Third, the NDR$_a$ assumes that consistency and regularity effects arise in a single-route lexical architecture. This stands in contrast to traditional theories that assume the necessity of a sub-lexical route to simulate these effects. Fourth, the leading dual-route model uses an interactive activation network in its lexical route, whereas the NDR$_a$ is build on the basis of discriminative learning principles. In what follows, we discuss each of these topics in more detail.

2.6.2 Serial versus parallel processing

The serial or non-serial nature of processing has been a central debate in the reading aloud literature (see M. Coltheart et al., 2001). Two types of experimental results are typically interpreted as evidence for serial processing. First, Weekes (1997) found a length by lexicality interaction, with a stronger effect of length in non-word reading than in word reading. Second, a number of studies (M. Coltheart & Rastle, 1994; Rastle & Coltheart, 1999; Roberts et al., 2003) found a position of irregularity effect with a larger processing cost when grapheme-to-phoneme irregularities occurred in early positions (e.g., *chef*) than when irregularities occurred in later positions (e.g., *blind*). These results have been taken as evidence for a dual-route architecture. In the dual route architectures of the CDP+ and DRC model the sub-lexical route operates in a serial manner: the uptake of orthographic information occurs in a letter-by-letter fashion.

The serial nature of the sub-lexical route is conceptually linked to a left-to-right moving window of spatial attention (Facoetti et al., 2006; Perry et al., 2007). By contrast, the lexical route of the CDP+ model processes the entire orthographic input at once and is therefore parallel in nature. In this framework, the interaction of length with lexicality results from the fact that non-word naming exclusively involves the serial sub-lexical route, whereas word naming also involves the parallel lexical route. In non-word naming additional letters lead to additional stages of information uptake and therefore longer naming latencies. This effect is diminished in word naming, because the parallel lexical route is insensitive to differences in word length (Perry et al., 2007).

Alternatively, length effects may be peripheral to the task of reading aloud and arise from extra-linguistic sources, such as processes related to articulation (Seidenberg & Plaut, 1998; Perry et al., 2007) or visual input decoding. In its current implementation, length effects in the NDR_a arise primarily as a result of visual input interpretation, which is consistent with an extra-linguistic interpretation of these effects. Nonetheless, the NDR_a correctly predicts that the length effect should be larger for words as compared to non-words. As noted by Perry et al. (2007), a potential source for the length by lexicality interaction in parallel models is dispersion. Non-words tend to have less common orthographic and phonological bigrams than real words. The larger length effect for non-words may therefore be a product of the increased likelihood of encountering a low frequency orthographic or phonological bigram in longer non-words. When a low frequency orthographic bigram occurs in a word, less activation is spread to orthographic neighbors, whereas when it contains a low frequency phonological bigram the activated neighbors will send less activation to the target demi-syllables. As such, low frequency orthographic and phonological bigrams both result in longer naming latencies.

While we believe that the length effect in word naming is at least partially driven by extra-linguistic processes, the non-serial nature of the NDR_a in its current form does not reflect a conceptual preference in the serial versus parallel processing debate. Indeed, the inability of the

current implementation of the NDR$_a$ to simulate the position of irregularity effect suggests that a serial uptake of information may be beneficial to the performance of the NDR$_a$ model. In a serial implementation, the position of irregularity effect would follow naturally from the increased availability of earlier orthographic input and phonological output units. Furthermore, Perry et al. (2007) demonstrated that the serialization of their sub-lexical route boosted item-level correlations significantly.

Sensitivity to the serial nature of the reading process already proved pivotal in the implementation of a verification mechanism for the pronunciation performance simulations. When selecting the first demi-syllable for pronunciation the checking mechanism ensures that only those phonological units are considered that are activated by semantic representations that share word-initial orthographic representations with the target word or non-word, whereas when selecting the second demi-syllable only those phonological units are considered that are activated by lexical representations that share word-final orthographic representations with the target word or non-word. This left-to-right nature of the checking mechanism at the lexical level showed improved performance over a similar parallel verification mechanism, particularly for the pronunciation of non-words.

2.6.3 Visual input interpretation

In the NDR$_a$, estimations of the time it takes to interpret the visual input are based on a rudimentary measure of the complexity of the visual input. When we developed the model we considered visual input interpretation peripheral to the linguistic core of the model and primarily implemented it as a convenient analogy to the feature detection systems in the DRC and CDP models. In our simulations, however, it became clear that the correct simulation of the length effect in the NDR$_a$ depends on the interpretation of the visual input. Given the importance of the length effect in the reading aloud literature, some further thought about the issue is warranted.

In its current form the visual input interpretation mechanism is insensitive to differences between words and non-words. Words and non-words alike are decomposed into letters and letter bigrams, which in turn ac-

tivate lexical representations. Evidence from the neuroscience literature, however, suggests that the early visual processing in occipital brain regions varies not only as a function of word length (Wydell et al., 2003; Tarkiainen et al., 1999; Indefrey et al., 1997), but also as a function of lexicality (e.g., Fiez et al., 1999; Xu et al., 2001, cf. Dehaene et al., 2002; Wydell et al., 2003 for studies that did not find lexicality-related differences of visual occipital region activations). Importantly, the visual occipital system is insensitive to linguistic properties of the input, which suggests that the observed effects of lexicality in this region reflect a difference in familiarity with the visual input across words and non-words.

A post-hoc analysis on the observed ELP naming latencies showed that a refinement of the visual input interpretation mechanism in the NDR_a that takes into account the familiarity of the visual input at the word level leads to a substantial improvement in item-level correlations. The predicted values of a simple linear model using as predictors the components of the NDR_a model, but replacing the complexity measure with a frequency-weighted alternative (i.e., *Complexity* divided by $log(Frequency)$ + backoff constant) showed a correlation of $r = 0.544$ to the observed naming latencies. Simply adding this frequency-weighted alternative to the NDR_a model, however, led to a poor qualitative performance of the model. Nonetheless, a visual input interpretation mechanism that takes into account the familiarity of the visual input in a more subtle manner, may well lead to further improvements in the performance of the NDR_a model. Such a visual complexity measure would fit well with the results of familiarization studies with objects and faces, in which greater occipital activation was found for unfamiliar objects and faces (Van Turennout et al., 2003; Rossion et al., 2003).

2.6.4 Orthographic input units

In the current implementation, orthographic representations in the NDR_a model are limited to letters and letter bigrams. Evidence from the neuroscience literature, however, suggests that this simple encoding scheme might be an oversimplification of the neurobiological reality of language processing. Vinckier et al. (2007) and Dehaene et al. (2005), for instance,

found that visual word recognition is sensitive to a hierarchy of increasingly complex neuronal detectors, ranging from letters to quadrigrams.

From a discrimination learning perspective the richness of the encoding scheme is an empirical issue. Language users extract those pieces of information from the input that provide valuable cues to the outcome. The current simulation results suggest that an encoding scheme based on letters and letter bigrams is sufficiently rich to capture a wide range of experimental findings in the reading aloud literature. If future experimental work indicates that higher order n-grams provide valuable additional information, however, we have no a priori objections against enriching the orthographic encoding scheme of the NDR$_a$. One possibility would be to include high frequency, but not low frequency letter n-grams as cues. Such a frequency-dependent coding scheme would help address the familiarity of the input issue raised above.

2.6.5 Phonological output representations

As for the orthographic input level, we also made a decision regarding the grain size of representations at the phonological output level of the NDR$_a$. At this level we decided to use demi-syllables (Klatt, 1979). The use of demi-syllables, however, is not free of problems. In its current implementation, for instance, the NDR$_a$ is not able to correctly simulate the reading aloud of non-words that contain non-existent demi-syllables. For instance, the predominant pronunciation of the non-word *filced* is *[fIlst]*, which includes the non-existing demi-syllable *[Ilst]*. Without a corresponding representation, the model cannot simulate the pronunciation of this demi-syllable.

Demi-syllables offered an easy-to-implement approximation of acoustic gestures that proved adequate for the current purposes. While this approximation worked well in the simulations reported here and shows that phoneme representation are superfluous for modeling reading aloud, we believe that an implementation of acoustic gestures at a finer grain size that more accurately reflects the biological reality of producing speech would further improve the performance of the NDR$_a$ and help develop an extension of the model to auditory language processing. One option

worth exploring in future research is the use of time-sensitive gestural scores as used in articulatory phonology (see, e.g., Browman & Goldstein, 1986, 1989, 1990, 1992).

2.6.6 Consistency effects in a lexical architecture

The effects of consistency and regularity have been important benchmark effects for models of reading aloud. The DRC model (M. Coltheart et al., 2001) successfully simulated the factorial effect of regularity (see, e.g., Seidenberg et al., 1994; Taraban & McClelland, 1987; Paap et al., 1987; Paap & Noel, 1991) through the grapheme-to-phoneme conversion rules in its sub-lexical route. These rules, however, operate in an all-or-none fashion. As a result, the DRC model did not capture graded consistency effects (Jared et al., 1990; Plaut et al., 1996; Jared, 1997; Rastle & Coltheart, 1999), which require the activation of not only the most common grapheme to phoneme mappings, but also that of other, less common mappings.

To overcome the difficulties of the DRC model, the CDP model uses the TLA sub-lexical network (Zorzi et al., 1998a; Zorzi, 1999) in its non-lexical route. As noted by Perry et al. (2007), the TLA sub-lexical network is a simple two-layer learning network that operates on the basis of the delta rule (Widrow & Hoff, 1960). One advantage of learning models over rule-based models is that they allow non-target words to influence the naming process (Treiman et al., 1995). As such, the TLA network allows for the successful simulation of graded consistency effects. In the CDP+ model the successful simulation of consistency effects, therefore, is a result of the associative learning in the sub-lexical route (Perry et al., 2007).

By contrast, M. Coltheart et al. (2001) suggested that consistency effects might arise in the lexical route as a result of neighborhood characteristics. Perry et al. (2007, p.276) contested this claim, stating that "such influences are too weak to account for the majority of the consistency effects reported in the literature". They support this claim by showing that consistency effects are still captured by a purely feedforward version of the CDP+ in which the activation of orthographic neighbors is

completely disabled. The fact that a sub-lexical network can generate consistency effects, however, does not provide conclusive evidence for the claim that a lexical network cannot.

To demonstrate this point we implemented a purely sub-lexical version of the NDR$_a$, in which orthographic units are mapped directly onto phonological outcomes. This sub-lexical version of the NDR$_a$ captured the linear effects of consistency ($t = -2.849$, $\beta = -0.063$), regularity ($t = -8.542$, $\beta = -0.375$) and friends minus enemies ($t = -2.382$, $\beta = -0.052$). The simulations with the original NDR$_a$ model, however, showed that all these effects can be captured in a lexical architecture as well. The fact that a sub-lexical network can capture the effects of consistency therefore does not imply that such a sub-lexical network is a necessary component of a model of reading aloud. The necessity for a sub-lexical route in the CDP+ model may not reflect the psychological reality of such a route, but instead display the shortcomings of the interaction activation model (McClelland & Rumelhart, 1981) that underlies the lexical route of the CDP+ model. We come back to this point in the next section.

In the lexical architecture of the NDR$_a$, regularity and consistency effects arise due to the co-activation of lexical items with similar orthographies. The co-activated lexical representations of consistent/regular words help co-activate the target word phonology, whereas the co-activated lexical representations of inconsistent/irregular words activate non-target phonological features. As a result, co-activated words help target word naming for consistent, but inconsistent words do not. In line with the suggestions of M. Coltheart et al. (2001), consistency effects in the NDR$_a$ therefore arise through neighborhood characteristics. These neighborhood characteristics did not only prove sufficient to simulate the observed effects of regularity and consistency in isolation, but also captured the complex interplay of these predictors as well as the interaction of consistency with friend-enemy measures (Jared, 1997, 2002) and frequency (Seidenberg et al., 1994; Taraban & McClelland, 1987; Paap et al., 1987; Paap & Noel, 1991). Furthermore, the NDR$_a$ captures the graded consistency effect for non-words (Glushko, 1979; Andrews & Scarratt, 1998). As such, the NDR$_a$ correctly simulates the complex and challenging pattern of results

for various orthography-to-phonology consistency measures through a purely lexical architecture.

2.6.7 Learning

The CDP+ model is a hybrid model that was built from a nested modeling perspective. The idea behind nested modeling is that a new model should be based on its predecessors (Jacobs & Grainger, 1994). Perry et al. (2007) therefore evaluated the strengths and weaknesses of the different components of the DRC and the CDP models. They found the rule-based sub-lexical route of the DRC model to be suboptimal and replaced it with the learning network of the CDP model (Zorzi et al., 1998b). On the other hand, the lexical route of the CDP model was not fully implemented and based on a simple frequency-weighted activation of a lexical phonology (Zorzi, 1999). They therefore replaced this with interactive activation network of the DRC model (M. Coltheart et al., 2001; McClelland & Rumelhart, 1981).

While we see the merit of a nested modeling approach, we are less convinced about the hybrid nature of the CDP+ model that resulted from it. Even if a dual-route model were conceptually correct one would expect that the lexical and non-lexical route operate on the basis of similar neuro-computational mechanisms. The implementation of the lexical route of the CDP+ seems particularly implausible given the fact that interactive activation models avoid the issue of learning (see, e.g., Baayen et al., 2011). Perry et al. (2007, p.303-304) acknowledge this problem and mention the lack of learning in the lexical route of the CDP+ model as one of its limitations. In addition, (Perry et al., 2007) state, the interactive activation model has been shown to fail to account for a number of findings in the lexical decision literature (see, e.g., Andrews, 1996; Ziegler & Perry, 1998). We therefore believe that a learning network implementation of the lexical route of the CDP+ model would be an option worth exploring.

A learning implementation of the lexical route would help establish the necessity for a dual-route architecture in the CDP+. In the current implementation of the CDP+ model the sub-lexical route has a substan-

tial independent contribution (Perry et al., 2007). This independent contribution, however, could have two sources. First, it could reflect the correctness of a dual-route architecture in which both routes reflect different parts of the language processing that occurs in the reading aloud task. Alternatively, however, the independent contribution of the sub-lexical route of the CDP+ model could be a result of the suboptimal performance of the interactive activation model in its lexical route, the contribution of which is currently limited to that of a frequency-weighted lexical phonology. In this case, the variance that is currently explained by the sub-lexical route of the CDP+ could also be explained by a better optimized lexical route. The finding that the addition of a non-lexical learning network did not improve the performance of the NDR_a supports such an interpretation.

2.7 Conclusions

We presented the NDR_a, a single-route model for reading aloud based on the fundamental principles of discriminative learning. The NDR_a is an extension of the NDR model by Baayen et al. (2011) for silent reading. We showed that the NDR_a provides a good overall fit to observed naming latencies. Through the use of generalized additive models we also demonstrated that the NDR_a successfully simulates not only the linear, but also the non-linear characteristics of a wide range of predictor effects and interactions documented in the experimental literature. As such, the NDR_a provides an alternative to leading models of reading aloud, such as the CDP+ (Perry et al., 2007) and CDP++ (Perry et al., 2010) models.

The NDR_a model is a major advancement over existing models of reading aloud in two ways. First, the computational engine of the NDR_a is based on the well-established learning algorithm provided by the Rescorla-Wagner (Rescorla & Wagner, 1972) equations. Given that the Rescorla-Wagner equations have been characterized as a general probabilistic learning mechanism (Chater et al., 2006; Hsu et al., 2010), the computational core of the model has increased biological plausibility

over models that assume language-specific processing mechanisms (see Baayen et al., 2011, 2013).

The learning architecture of the NDR$_a$ stands in contrast to the lexical route of the CDP+ model, which is based on the interactive activation model of McClelland and Rumelhart (1981). In the current implementation of the CDP+ model the contribution of the lexical route is "limited to the provision of frequency-weighted lexical phonology" (Perry et al., 2007, p.303). Perry et al. (2007, p.303-304) acknowledge the problems associated with the interactive activation model in their lexical route and name the lack of learning in the lexical route of the CDP+ as one of its shortcomings.

The discriminative learning mechanism underlying the NDR$_a$ also differs substantially from the connectionist principles that form the computational basis of the different versions of the triangle model (see, e.g., Seidenberg & McClelland, 1989; Harm & Seidenberg, 1999; Seidenberg & Gonnerman, 2000; Harm & Seidenberg, 2004). As noted by Baayen et al. (2013), the computational engine of the NDR$_a$ is much simpler than that of connectionist models. The NDR$_a$ learning networks directly map input units onto outcomes, without the intervention of one or more layers of hidden units.[7] The NDR$_a$ is therefore more transparent than connectionist models, with activations of output units representing simple posterior probability estimates of outcomes given input units. In addition, in contrast to connectionist models the NDR$_a$ does not rely on the neurobiologically implausible process of back-propagation learning.

The second major advancement of the NDR$_a$ is that it uses a single lexical route architecture for both word and non-word naming. We showed that a single lexical route based on discriminative learning not only provided a good overall fit to observed naming latencies, but also captured a number of experimental results that are typically attributed to processes in the sub-lexical route. The non-linear main effects and interactions of consistency and regularity measures, for instance, are accurately captured

[7] Note that the latest version of the triangle model does not contain hidden layer units, but, instead, operates on the basis of a direct mapping between input units and outcomes.(Harm & Seidenberg, 2004)

by the NDR$_a$. In addition, we showed that the NDR$_a$ makes predictions for non-word naming that are highly similar to those of the dual-route CDP+ model. Furthermore, we documented the existence of a strong non-word frequency effect in the classic McCann and Besner (1987) non-word reading latencies, which provides evidence for the involvement of a lexical route architecture in non-word naming.

The single-route architecture stands in contrast to the dual-route architectures of leading models of reading aloud, including both traditional dual-route models such as the DRC (M. Coltheart et al., 2001; McClelland & Rumelhart, 1981), CDP (Zorzi et al., 1998b), CDP+ (Perry et al., 2007) and CDP++ (Perry et al., 2010) and the most recent versions of the triangle model (see, e.g., Harm & Seidenberg, 2004). These models contain both a direct orthography to phonology mapping and an orthography to phonology route that is mediated by semantics. While the non-lexical route of the CDP+ model has a significant contribution to the model performance (see Perry et al., 2007), we demonstrated that the addition of a non-lexical discriminative learning network does not improve the performance of the NDR$_a$ model.

The current implementation of the NDR$_a$, however, provides a highly simplified window on reading aloud. At both the orthographic and the phonological level we make use of discrete representations at a highly restricted subset of possible grain sizes. Findings from the neuroscience literature (see, e.g., Vinckier et al., 2007; Dehaene et al., 2005) suggest that a more flexible system operating over multiple grain sizes may further improve the performance of the model.

In addition, the current simulations focused on the unimpaired language processing system. A substantial amount of work has been carried out on impaired language processing in both surface and deep dyslexia patients (see e.g., Patterson & Behrmann, 1997; Derouesne & Beauvois, 1985). It will be interesting to see to what extent selective lesioning of the discriminative learning networks could capture the patterns of results seen in these patients. One possibility is that the prefrontal structures and conflict resolution skills that underlie target pronunciation selection in the NDR$_a$ may not be as easily accessible when the system is lesioned,

possibly due to capacity limitations. Such an interpretation would fit well with the findings of Hendriks and Kolk (1997), who demonstrated that the behavioral symptoms used to classify dyslexic patients into deep and surface dyslexia arise not only as a result of deficiencies in the language processing system, but also due to strategic choices in the context of the task at hand.

Furthermore, similar to the CDP+ model, the current implementation of the NDR_a processes mono-syllabic words only. Perry et al. (2010) extended the CDP+ to allow for the processing of both mono- and bi-syllabic words. The extension of the NDR_a to reading beyond the single syllable level is a further topic to explore in future research.

In its current state, however, the NDR_a provides a single-route alternative to state-of-the-art dual route models of reading aloud that is based on a simple general learning algorithm and that - with a parsimonious architecture - accurately captures many of the linear and non-linear patterns in experimental word and non-word reading data.

3

Compound reading

3.1 Introduction

Ever since Rayner (1978), psycholinguists have been using eye fixation patterns as a proxy for language processing costs during reading. While traditional behavioral measures obtained from tasks such as lexical decision or reading aloud generate unidimensional measures of the end-point of processing, eye-tracking offers an on-line measure of linguistic processing with a high temporal resolution. As such, the eye-tracking methodology offers the opportunity to investigate how language processing unfolds over time.

Large-scale databases for traditional behavioral measures offer a wealth of information. The English Lexicon Project (Balota et al., 2007), for instance, contains more than $2,700,000$ reaction times for lexical decision and over $1,100,000$ naming latencies. Resources for eye-tracking, however, are much more limited. The Potsdam Sentence Corpus (Kliegl et al., 2006) contains around $150,000$ fixations for 254 participants in a sentence-reading task in German. The Dundee Corpus (Kennedy, 2003) is the largest available corpus of eye-tracking fixations and contains over $400,000$ fixations per language on newspaper articles in both English and French.

In this chapter we use a new large-scale eye-tracking database: the Edmonton-Tübingen eye-tracking corpus (henceforth ET corpus). For the ET corpus we recorded the eye movements of 6 participants as they read a collection of fictional texts, consisting of the fiction section of the Brown University Standard Corpus of Present-Day American English (henceforth Brown Corpus; Francis & Kucera, 1979) as well as fictional texts by Lewis Carroll (Alice in Wonderland and Through the Looking Glass) and Sir Arthur Conan Doyle (The Adventures of Sherlock Holmes). For ecological validity, all texts in the ET corpus experiment were presented as full pages of text. The ET corpus contains 100 hours of eye movement data per participant, for a total of 2,019,997 eye fixations.

The scale of the ET eye-tracking corpus enables us to investigate language processing during reading with a high temporal resolution, even when the frequency of occurrence of the linguistic phenomenon of interest is relatively low. Here, we use the richness of the ET corpus data to investigate morphological processing of noun-noun compounds in a naturalistic reading task.

Compound processing has received considerable attention in the experimental psycholinguistic literature. A host of behavioral experiments investigating how morphologically complex words are processed has been carried out, including lexical decision experiments (see, e.g., Taft & Forster, 1975, 1976; Taft, 1979; Van Jaarsveld & Rattink, 1988; De Jong et al., 2002; Duñabeitia et al., 2007; Marelli & Luzzatti, 2012), word naming experiments (see, e.g., Juhasz et al., 2003; Baayen et al., 2010), priming studies (see, e.g., Monsell, 1985; Zwitserlood, 1994; Jarema et al., 1999; Libben et al., 2003) and masked priming studies (see, e.g; Giraudo & Grainger, 2001).

More recently, the eye-tracking methodology has been used to provide further insight into the time course of compound processing. While some studies presented compounds in isolation (see, e.g., Kuperman et al., 2009), most eye-tracking experiments embedded compounds in sentence contexts in an attempt to improve the ecological validity of the data (see, e.g., Hyönä & Pollatsek, 1998; Pollatsek et al., 2000; Bertram & Hyönä,

2003; Juhasz et al., 2003; Andrews et al., 2004; Pollatsek & Hyönä, 2005; Kuperman et al., 2008).

In this chapter, we look at the eye fixation patterns for noun-noun compounds in the ET corpus. Using the ET corpus data offers increased ecological validity over experiments that present compounds in isolated sentence contexts. First, while sentence contexts provide a local context for the embedded compounds, normal discourse structure is disrupted on a trial-by-trial basis in these experiments. By contrast, the compounds in the ET corpus appear in a sensibly developing flow of text. This allows readers to form better-informed hypotheses about the upcoming linguistic information, which could lead to substantial differences in the way compounds are processed.

Second, the within-experiment frequency of compounds is, by definition, artificially high in experiments that are designed to investigate compound processing. Even when compounds are embedded in sentences and filler items are added to the experimental lists, participants may consciously or subconsciously be aware of the increased prevalence of morphologically complex words. By contrast, the frequency of occurrence of compounds in the ET corpus better reflects the frequency of compounds in the language as a whole. Unlike in compound experiments, therefore, participants in the ET corpus experiment are unlikely to attribute attentional resources to compounds that would not be recruited for compound processing in everyday language use. As such, the ET corpus data offer the possibility to look at compound processing in a highly naturalistic linguistic environment.

Previous experimental work has led to a wide range of psycholinguistic theories about compound processing (see Kuperman et al., 2009 and Kuperman et al., 2008 for a comprehensive overview of compound processing theories). According to sub-lexical models, compounds are decomposed into the constituent morphemes and subsequently the full form of the compound is accessed through these constituent morphemes (see, e.g., Taft & Forster, 1975, 1976; Taft, 1979, 1991, 2004). Sub-lexical models differ with respect to the relative importance they attribute to the constituent morphemes. Some sub-lexical models deem the left constitu-

ent pivotal for accessing the full form of the compound (Taft & Forster, 1976), while other sub-lexical models consider the right constituent essential for full-form access (Juhasz et al., 2003). In contrast to sub-lexical theories, supra-lexical theories propose that full-form access temporally precedes access to the constituent morphemes of a compound (see, e.g., Giraudo & Grainger, 2001; Diependaele et al., 2005).

While sub-lexical and supra-lexical theories assume that the activation of a compound and its constituents is sequential in nature, parallel dual-route models argue for the simultaneous activation of a compound and its constituents. Most parallel dual-route models propose a horse race between the full-form route and the decompositional route (see, e.g., Schreuder & Baayen, 1995; Baayen & Schreuder, 1999; Allen & Badecker, 2002), but dual-route models in which both routes are allowed to interact exist as well (see, e.g., Baayen & Schreuder, 2000).

Finally, maximization of opportunity theories of compound processing propose that readers use simultaneously all available information and all available processing mechanisms to process compounds (Libben, 2005, 2006). On the basis of recent experimental findings, for instance, Kuperman et al. (2009, p.887) concluded that:

> Effectively, a model that meets these requirements is no longer a dual route model, but rather a multiple route model that, in morphological terms, allows access to full-forms, immediate constituents, embedded morphemes and morphological families. More generally, such a model will have as its basic principle maximization of all opportunities, both morphological, orthographic, phonological, and contextual, for comprehension of the visual input. We believe that probabilistic and information-theoretical approaches to lexical processing developed recently in morphological and syntactic research [...] hold promise for formalization of those opportunities and for computational implementation of the multiple-route model of compound recognition.

A full-fledged maximization of opportunity model that uses all orthographic, morphological, phonological and contextual information available to the reader presents a non-trivial computational challenge. Naive Discrimination Learning (henceforth NDL) offers a simpler computational implementation of a functional model of compound processing that allows for the simultaneous activation and interaction of a large number of lexemes, including the lexemic representations of a compound and its constituents. Baayen et al. (2011) demonstrated the explanatory power of the NDL framework for compound processing by successfully modeling the effects of compound length, compound frequency, modifier frequency and modifier family size in the lexical decision latencies for a large set of compounds in the English Lexicon Project (Balota et al., 2007).

Here, we investigate the explanatory power of standard lexical predictors, as well as that of systemic measures derived from orthography-to-lexeme and lexeme-to-lexeme NDL networks, for the eye fixation patterns in the ET corpus. Lexical predictors provide insight into which lexical properties of words correlate with behavioral measures of language processing. A lexical predictor analysis, however, provides little insight into *why* certain lexical predictors show a correlation with behavioral measures, whereas others do not. By contrast, NDL measures are directly derived from discrimination learning networks. As a result, these measures have the potential to inform us about the properties of the learning system that drive patterns of results for behavioral measures.

Discrimination learning has been demonstrated to successfully account for a wide range of experimental effects in lexical decision studies (see, e.g., Baayen et al., 2011, 2013; Ramscar et al., 2014), as well as for word naming (see Chapter 2). The current study, however, is the first to investigate to what extent the NDL model provides a systemic alternative to standard lexical predictors for eye movement data.

There is one important difference between the simulations of lexical decision latencies and word naming latencies on the one hand and the simulation of eye movement patterns on the other hand. For lexical decision and word naming, the studies mentioned above aimed at directly modeling the response variable: reaction times. The information uptake

process in natural discourse reading, however, is much more complex than that in single word lexical decision or word naming studies. Eye movement patterns are determined not only by lexical properties of the linguistic input, but also by a host of visuomotor processes. The aim of the current study, therefore, is not to generate simulated values for a response variable, but to investigate to which extent systemic predictors derived from discrimination learning networks can provide further insight into the processes that underlie compound reading in natural discourse.

3.2 Methods

3.2.1 Participants

The data of the four participants in the Edmonton-Tübingen eye-tracking corpus (henceforth ET corpus) for which data preprocessing has been completed were used. All participants were graduate students at the University of Alberta. Two participants were male, two were female. Both male participants and one female participant were native speakers of English, whereas the other female participant was a near-native speaker of English. The male participants were 32 and 27 years old, the female participants were 29 and 25 years old. All participants had normal or correct-to-normal vision. Participants received $20 per hour for their participation.

3.2.2 Materials

For present purposes, we used the subset of the ET corpus that consisted of the eye movement data for the fiction section of the Brown Corpus (Francis & Kucera, 1979). The fiction section of the Brown Corpus consists of 126 samples from fictional texts, subdivided into 6 genres: general fiction (29 texts), mystery and detective (24 texts), science fiction (6 texts), adventure and western (29 texts), romance and love stories (29 texts) and humor (9 texts). The 126 fictional text samples in the Brown Corpus consist of 253, 092 word tokens. In total, the 4 participants

included here fixated on these 253,092 word tokens 923,659 times, for an average of 230,915 fixations per participant.

Here, we are interested in the set of noun-noun compounds that appear in this fiction section of the Brown Corpus. On the basis of the CELEX lexical database (Baayen et al., 1995) we therefore compiled a list of noun-noun compounds and their constituents. From the data we then extracted all singular and plural forms of noun-noun compounds that were present in this list of noun-noun compounds. We initially included not only compounds that appeared as such in the data (e.g., "airplane"), but also space-separated (e.g., "home plate") and hyphen-separated (e.g., "turtle-neck") forms of these compounds. In total, this procedure yielded 49 space-separated noun-noun compound types (11.34%), 19 hyphen-separated compound types (4.40%) and 364 non-separated noun-noun compound types (84.26%). Given the small proportion of hyphen-separated and space-separated compound types, we decided to limit the data set to non-separated compounds. The remaining 364 compound types correspond to 844 compound tokens in the Brown Corpus fiction section. In total, the 4 participants used here fixated 4747 times on these 844 compound tokens.

3.2.3 Design

3.2.3.1 Response variables. For each fixation on a noun-noun compound, we extracted the fixation duration, the fixation position and whether or not the compound was fixated on again. Fixation duration is the duration in milliseconds of a given fixation of the eye. Fixation durations were log-transformed to remove a rightward skew from the fixation duration distribution. Outlier fixations shorter than 60 milliseconds (2.42%) or longer than 500 milliseconds (0.34%) were removed from the data prior to analysis.

Fixation position is the position of a fixation, measured in number of pixels from the left boundary of the word. A visual inspection of the fixation position data indicated that there was a substantial amount of vertical drift in the data, particularly near the end of lines. To ensure that the interest area associated with a fixation was reliable, a correction

algorithm was used to correct for this vertical drift. The performance of the correction algorithm was visually inspected for each fixation and fixation positions were corrected manually when necessary. Fixation positions further than 2.5 standard deviations from the fixation position mean were removed from the data prior to analysis (1.60%).

The final response variable considered in this study is a binary variable that encodes whether or not the compound was fixated on again after the current fixation. If no additional fixations were necessary this indicates that processing was completed to a sufficient extent to proceed to the next word. By contrast, additional fixations indicate that readers were unable to complete compound processing during the current fixation. Thus, the probability of a refixation gauges how successful processing during the current fixation was.

3.2.3.2 Experimental predictors. For each fixation on a noun-noun compound four control predictors were extracted from the ET corpus data: the experimental session (Session), the page number within an experimental session (Page), the line number within a page (Line) and the horizontal position of a fixation on the page (X Page).

We furthermore more looked at the incoming saccade length. Due to the full text reading task used here, however, the distribution of incoming saccade lengths was strongly bimodal. This bimodal distribution was characterized by a large peak at small to medium positive values for normal rightward saccades and a smaller peak at large negative values for saccades that involved moving the eye to the start of a new line. The bimodal distribution of incoming saccade length was resistant to normalization procedures. We therefore decided not to include saccade length in statistical models reported below. We did, however, run post-hoc analyses to verify that the results reported here remain significant when saccade length is entered into the model. Whenever an effect no longer reaches significance when saccade length is entered into the model, this is reported in the discussion of that effect.

3.2.3.3 Lexical predictors. Nineteen lexical predictors were included in the design, all of which were scaled prior to analysis. Whenever a predictor was included in a statistical model, all predictor values further than 2.5 standard deviations from the mean were removed prior to analysis to prevent artifactual outlier effects in the GAMMs fitted to the data.

For each compound, the length in letters of the modifier (Modifier Length), the head (Head Length) and the compound (Compound Length) were included as lexical predictors. Furthermore, the frequency of the modifier (Modifier Frequency), the head (Head Frequency) and the compound as a whole (Compound Frequency) in the British National Corpus (henceforth BNC; Burnard, 1995) were included as predictors. Compound frequencies were token frequencies of the inflectional variant (as it appeared in the Brown Corpus), rather than lemma frequencies (e.g., the frequency of "snowflakes" was that of the form "snowflakes", rather than the summed frequency of the singular form "snowflake", the plural "snowflakes" and the genitive forms "snowflake's" and "snowflakes'" used by for instance Kuperman et al., 2009). We log-transformed all frequency measures to remove a rightward skew from the frequency distributions.

In addition to the frequencies of the modifier, head and compound, the average letter bigram frequency for each of these components was included as a lexical predictor. The resulting predictors Modifier Mean Bigram Frequency, Head Mean Bigram Frequency and Compound Mean Bigram Frequency were extracted from the English Lexicon Project (henceforth ELP; Balota et al., 2007). From the ELP we also obtained two measures of orthographic neighborhood density: N-Count (the number of orthographic neighbors (M. Coltheart et al., 1977) and Orthographic Levenshtein Distance 20 (henceforth OLD; the average string edit distance between a word and its 20 closest neighbors (Yarkoni et al., 2008). Applied to the modifier, the head and the compound as a whole this resulted in five additional lexical predictors: Modifier N-Count, Modifier OLD, Head N-Count, Head OLD, and Compound OLD. Compound N-Count was not included in the analysis, because a large majority of compounds tokens had 0 orthographic neighbors. The OLD measures

were log-transformed prior to analysis to remove a rightward skew from the OLD distributions.

Morphological family size is a measure of the number of compounds both constituents occur in. For a given compound, the family size of the modifier is defined as the number of compounds that have the same left constituent as that compound. Similarly, the family size of the head is defined as the number of compounds that have the same right constituent as a given compound. For the compound "starlight", for instance, "stardust" and "starfish" are members of the morphological family of the modifier, whereas "daylight" and "skylight" are members of the morphological family of the head. The morphological family sizes of the modifier and the head have been shown to influence compound processing in a variety of tasks (see, e.g., De Jong et al., 2000, 2002; Dijkstra et al., 2005; Moscoso del Prado Martín et al., 2004; Juhasz & Berkowitz, 2011). We therefore added the Modifier Family Size and Head Family Size as calculated from the CELEX lexical database to the set of lexical predictors. Both family size measures were log-transformed prior to analysis to remove a rightward skew from the family size distributions.

The final three lexical predictors concern the semantic similarity of the modifier, the head and the compound as a whole, as gauged through Latent Semantic Analysis (henceforth LSA) similarity scores (see Landauer et al., 1998). LSA Similarity Modifier-Head is the term-to-term LSA similarity of the modifier and the head, whereas LSA Similarity Modifier-Compound and LSA Similarity Head-Compound refer to the LSA similarity of the compound to its modifier and head.

3.2.3.4 NDL predictors. To investigate the systemic correlates of compound processing we trained two NDL networks on the 102, 268, 226 words of the British National Corpus (Burnard, 1995). The first NDL network maps orthographic features (cues) onto lexemes (outcomes). Bearing in mind the maximization of opportunities hypothesis put forward by Libben (2006), we followed Baayen et al. (2011) in allowing for contextual learning by using not only the orthographic trigrams of the target word,

but also the orthographic trigrams of the preceding two words as input cues for the target lexeme.

Each three-word sequence in the British National Corpus (henceforth BNC) was presented as a learning event to the orthography-to-lexeme discrimination learning network. A learning event consists of the presentation of a set of input cues and the subsequent adjustment of the network weights between these input cues and all outcomes based on the discrepancy between the expected activation and the actual activation of each outcome unit. The orthographic trigrams in a three-word sequence were the cues for a learning event, while the lexemes corresponding to each of the three words were the outcomes. The character "#" was used to encode a spatial separation between subsequent cues. For the three-word sequence "that delicious cocktail", for instance, the cues would be "#th", "tha", "hat", "at#", "t#d", "#de", "del", "eli", "ici", "cio", "iou", "ous", "us#", "s#c", "#co", "coc", "ock", "ckt", "kta", "tai", "ail" and "il#", whereas the outcomes would be "THAT", "DELICIOUS" and "COCKTAIL".

Baayen et al. (2011) proposed a full-decomposition model of morphological processing in which no lexemic representations for morphologically complex words exist. While such an approach works well for inflected words and transparent compounds, opaque compounds such as "cocktail" pose a challenge to this modeling strategy. We therefore use a holistic modeling strategy, in which a separate lexemic representation is posited for each compound. The outcome associated with a training event in which the orthographic cues of a compound were encountered was this compound lexeme, rather than the lexemes of the constituents of the compound (see also Pham & Baayen, 2015, for a similar modeling strategy for compounds in Vietnamese).

From a learning perspective a holistic training regime for compounds makes sense. For each learning event, the weights from the set of cues in the input to all outcomes are updated. Each occurrence of the word "cocktail", for instance, reinforces the connections between the orthographic cues in the word "cocktail" and the lexeme COCKTAIL. At the same time, the absence of the outcomes COCK and TAIL when the orthographic features of the word "cocktail" are present in the input allows the model

to dissociate the lexemes COCK and TAIL from the orthographic cues in "cocktail". In other words: using a holistic training regime allows the model to learn that the whole may be more than, and often quite different from, the sum of the parts.

Importantly, the holistic training regime adopted here does not imply that the lexemes of the compound constituents are not activated in a bottom-up fashion when the orthographic form of a compound is encountered. Quite the contrary: due to the orthographic overlap between the compound as a whole and its constituents, the constituent lexemes generally receive a considerable amount of activation when the target word is presented – even in a network that is trained holistically. When we refer to the training regime used here as holistic we refer to the holistic nature of the top-down adjustment of weights in the Rescorla-Wagner network during learning, not to the bottom-up activation of lexemes.

The orthography-to-lexeme network resulted in 5 systemic estimates of processing costs, all of which were straightforwardly derived from Equation 1.3. First, the activation of the modifier given the orthographic trigrams of the compound as a whole (NDL Activation Modifier), as well as the activation of the modifier given the orthographic trigrams of the modifier itself (NDL Self-Activation Modifier) were included as NDL predictors. NDL Activation Modifier measures the degree to which the orthographic presentation of the compound activates the lexeme of the modifier, whereas NDL Self-Activation Modifier looks at the activation of the modifier lexeme when only the modifier is present in the orthographic input. The correlation between NDL Activation Modifier and NDL Self-Activation Modifier was $r = 0.414$.

Similarly, the activation of the head given the orthographic features of the compound as a whole (NDL Activation Head), as well as the activation of the head given the orthographic features of the head itself (NDL Self-Activation Head) were extracted from the orthography-to-lexeme NDL network. The correlation between NDL Activation Head and NDL Self-Activation Head was $r = 0.582$.

In addition to the activation of the modifier and the head, we furthermore extracted the activation of the compound lexeme given the orthographic features of the compound. This simple measure of the bottom-up support for the compound lexeme performed quite well. We found, however, that a richer measure that integrates the bottom-up support for the compound lexeme, the head lexeme and the modifier lexeme in an additive fashion provided maximum explanatory power. We therefore defined NDL Activation Compound as a weighted sum of the activation of the compound, the head and the modifier lexemes given the orthographic trigrams of the compound (see Baayen et al., 2011 for further examples of such a weighted integration):

$$\text{NDL Activation Compound} = a_{modifier} + a_{head} + w_1 * a_{compound} \quad (3.1)$$

with $a_{modifier}$, a_{head} and $a_{compound}$ being the activation of the modifier, head and compound lexemes given the orthographic trigrams of the compound and w_1 being a weight parameter for the relative contribution of the compound lexeme as compared to the head and modifier lexemes. For all analyses reported below, w_1 was set to 1.2.

The fact that an additive integration of the compound, head and modifier lexeme activations proved optimal demonstrates that even when using a holistic training scheme, the degree to which the orthographic form of a compound activates not only the compound lexeme, but also the constituent lexemes helps explain variance in eye movement data. For English, a weighted additive integration of the compound lexeme activation and the constituent lexeme activations suffices. As we shall document below, a greater activation of the constituent lexemes leads to faster and more successful compound processing. Importantly, this is a consequence of the distributional properties of the English language, rather than a general property of morphological processing. For Vietnamese, for instance, Pham and Baayen (2015) showed that greater activation of the compound constituents leads to longer, rather than shorter processing times.

The distributions of the NDL activation predictors described above were characterized by a rightward skew. To remove this rightward skew, we log-transformed all NDL activation predictors. Furthermore, a backoff constant of 0.05 was added to all activations prior to this log-transform to prevent taking the logarithm of a non-positive number.

The second NDL network used in the current simulation learns association strengths between lexemes and lexemes. For each of the 102,268,224 word trigrams in the British National Corpus, we used the lexeme representations of the first two words as cues and the lexeme of the third word as outcome. For the trigram "that delicious cocktail", for instance, the cues would be "THAT" and "DELICIOUS", whereas the outcome would be "COCKTAIL". The second NDL network reflects contextual learning at the lexeme level. As before, training was entirely holistic in nature, with independent lexemic representations for each compound.

Four NDL predictors were derived from the lexeme-to-lexeme NDL network. The first three predictors can be regarded as the NDL counterparts of the LSA lexical predictors and gauge the semantic similarity of the head, the modifier and the compound as a whole. From the lexeme-to-lexeme network we extracted the column vectors of weights for the head, the modifier and the compound as a whole given all word types in the training lexicon. We then calculated the correlations between these vectors to obtain the similarity of the three components. NDL Similarity Head-Modifier is the correlation between the weight vectors of the head and the modifier, NDL Similarity Head-Compound the correlation between the weight vectors of the head and the compound and NDL Similarity Modifier-Compound the correlation between the weight vectors of the modifier and the compound. To normalize the distributions of the NDL similarity measures, we inverse-transformed each similarity measure ($f(x) = -\frac{1}{x}$).

The fourth NDL predictor extracted from the lexeme-to-lexeme NDL network is the median absolute deviation (henceforth MAD) of the vector of compound weights given all word types in the training lexicon. The median absolute deviation is a measure of dispersion that is more robust to outliers than the standard deviation. Conceptually, it can be thought of

as a measure of network connectivity: the higher the MAD of a compound, the greater its network connectivity and the easier it is to access its lexeme. In other words, the NDL MAD measures provide a systemic measure of the prior probability of a lexeme (see Milin et al., 2015 for an application of the MAD measure in the context of discrimination learning). We log-transformed NDL MAD compound to remove a rightward skew from its distribution.

In total, the orthography-to-lexeme (5 predictors) and lexeme-to-lexeme (4 predictors) networks resulted in 9 systemic measures of lexical processing. All NDL measures were scaled. Predictor values further than 2.5 standard deviations from the predictor mean were considered outliers and were removed from the data prior to analysis whenever the relevant predictor was included in a statistical model.

For each response variable, we jointly removed outliers for the relevant lexical predictors and NDL measures to obtain identical data sets for the lexical predictor and NDL analyses. This allowed us to directly compare the goodness-of-fit of both models. As a result of this procedure, 8.31% of the data for the first-and-only fixation duration models, 2.15% of the data for the probability of refixation models, 7.03% of the data for the first-of-many fixation duration models, 1.82% of the data for the second fixation duration models, 2.09% of the data for the second fixation position models, 2.10% of the data for the probability of third fixation models, 8.27% of the data for the third fixation duration models and 0.79% of the data for the third fixation position models were excluded prior to analysis. No data points were excluded for the first-and-only fixation position and first-of-many fixation position models.

3.2.4 Procedure

Eye movements in the ET corpus were recorded with an EyeLink 1000 system using a temporal resolution of 500 Hz. Stimuli were presented on a 17-inch CRT monitor using a 1024 by 768 pixel resolution. Participants read with the head positioned on a chin rest that was located at a distance of 70 cm from the monitor. Prior to the experiment, participants were trained for 1 hour to self-calibrate with a game pad using the 9 point

calibration method. Participants were instructed to read at a natural pace, but to limit eye blinking to a minimum throughout the experiment.

A fixation mark that was used for drift checking was shown prior to each page of text at the location of the first letter of the text. Fixating on the fixation mark triggered the presentation of a page of text. Participants were instructed to read the page and press a button on a game pad to move on to the next page of text. A 9-point self-calibration was carried out after every 5 pages. Each experimental session consisted of a minimum of 30 pages of text and a maximum of 43 pages of text (mean: 36.29, sd: 2.52), presented in black 26 point Courier New Bold font against a white background.

The total reading time for each participant in the ET corpus was 100 hours. For present purposes, we limited the data to the 63 hour subset of the data from the fiction section of the Brown Corpus. This subset of the data consisted of 126 experimental sessions. Each experimental session had a duration of about 30 minutes, including a 5 minute break in between sessions. Participants were instructed to take a 10 minute break after each hour and to run no more than 4 sessions at a time.

3.3 Analysis

We fitted separate statistical models for the fixation position and fixation duration of first-and-only fixations, first-of-many fixations, second fixations and third fixations. Furthermore, we analyzed the probability of refixation for the first fixation and second fixation data. We also investigated the fixation duration and position of fixations preceding the first fixation of a compound, but found no statistically robust effects of lexical predictors or NDL measures on these response variables.

For each subset of the data, we fitted a statistical model using standard lexical predictors and a model using NDL measures. Typically, the results of lexical predictor models are used to inform the architecture of interactive activation models of language processing similar to McClelland and Rumelhart (1981), with effects of linguistic predictors being interpreted as evidence for mental representations related to the predictor(s)

in question (see Baayen et al., 2013). The statistical models using standard lexical predictors thus are a proxy of the potential of interactive activation accounts of compound processing, be they sub-lexical, supra-lexical or dual-route models. By contrast, the statistical models using NDL measures gauge to what extent the systemic approach proposed here can provide further insight into the eye fixation patterns in compound reading.

All analyses were carried out using generalized additive mixed-effect models (henceforth GAMMs, (see Hastie & Tibshirani, 1986), as implemented in version 1.8-7 of the *mgcv* package for R (Wood, 2006, 2011). Standard GAMMs were used to model the fixation duration and fixation position data for first-and-only fixations, first-of-many fixations, second fixations and third fixations. Logistic GAMMs were used to model the probability of refixation for first and second fixations. By-participant and by-item random intercepts were included when significant. Whenever necessary, main effect smooths for lexical predictors and NDL predictors were limited to 6 knots to prevent uninterpretable and in all likelihood over-fitted effects. The relative performance of the lexical predictor models and the NDL models was evaluated by comparing the restricted maximum likelihood (henceforth REML) scores (standard GAMMs) or unbiased risk estimator (henceforth UBRE) scores (logistic GAMMs) of both models. For interpretability, visual presentations of main effect smooths represent the partial effect of a predictor plus the model intercept.

3.4 Results

3.4.1 Single fixation duration

Of all 3117 first fixations in the ET corpus compound data, 1911 fixations were first-and-only fixations. In other words: 61.31% of all compounds were only fixated on once. By comparison, a mere 18% of Dutch compounds presented in isolation in Kuperman et al. (2008) required a single fixation only, whereas only 28% of the (tri-morphemic) Finnish compounds presented in sentential contexts in Kuperman et al. (2009) were

processed during a single fixation. The high proportion of single fixations in the current data demonstrates that a majority of compounds can be processed during a single fixation when compounds are embedded in natural discourse contexts.

3.4.1.1 Lexical predictor model. To investigate the lexical properties that influence the duration of these first-and-only fixations, we fitted a GAMM with random effect terms for Participant ($F = 11.812, p < 0.001$) and Word ($F = 0.113, p = 0.039$) to the fixation durations. We found main effects of Line ($F = 9.341, p < 0.001$), X Page ($F = 14.267, p < 0.001$) and X Word ($F = 3.236, p = 0.030$). These main effects are visualized in Figure 3.1. The left panel of Figure 3.1 presents the effect of Line and shows that fixation durations become longer as the vertical position of a compound on the page increases (i.e., as fixations are further down the page). By contrast, the middle panel of Figure 3.1 shows that fixation durations become shorter as the horizontal position of a compound on the page increases. This effect of X Page suggests that more preceding contextual information on the same line allows for faster processing. The effect of X Word (i.e., the fixation position within a compound) plotted in the right panel of Figure 3.1 indicates that more rightward fixations on the compound lead to shorter fixation durations.

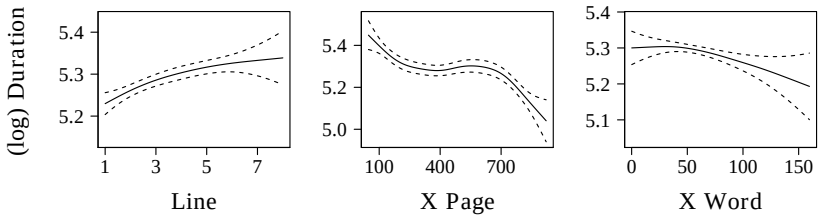

Figure 3.1. Effects of Line (left panel), X Page (middle panel) and X Word (right panel) on (log) Fixation Duration of first-and-only fixations.

In addition to the main effects of Line, X Page and X Word, we observed an interaction between Compound Frequency and LSA Similarity Compound Head ($F = 3.245, p = 0.003$). This interaction is presented in Figure 3.2. The z-axis shows predicted fixation durations, with warmer colors representing longer fixation durations and colder colors representing shorter fixation durations. The early effect of the semantic similarity of a compound and its head is in line with the early semantic effect observed by Marelli and Luzzatti (2012). Fixation durations are shorter for compounds that are semantically similar to their heads. As can be seen in Figure 3.2, the effect of LSA Similarity Head-Compound is very pronounced for low-frequency compounds, but much less prominent for high-frequency compounds. This indicates that category congruence may be important for low frequency compounds, but is much less relevant for high frequency compounds. For a frequent compound like "nightmare", the fact that a "nightmare" is not a female horse has little to no impact on amount of time required for successful processing. For an infrequent compound like "sagebrush", however, the semantic incongruence between the compound and its head may slow down processing considerably.

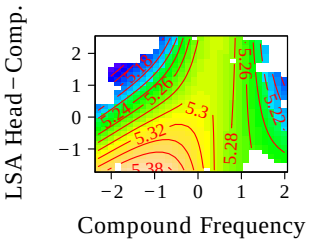

Figure 3.2. Effect of the interaction between Compound Frequency and LSA Similarity Head-Compound on (log) Fixation Duration of first-and-only fixations.

For all but the lowest values of LSA Similarity Compound-Head, the effect of Compound Frequency on first-and-only fixation durations is inverse U-shaped in nature, with shorter fixation durations for compounds with non-typical frequencies. Previously, a number of studies found weak non-significant effects of compound frequency on first fixation duration

(see, e.g., Zwitserlood, 1994; Bertram & Hyönä, 2003; Andrews et al., 2004). More recently, Kuperman et al. (2009) were the first to observe a robust significant effect of compound frequency on first fixation patterns. The current effect of compound frequency provides further support for the early emergence of full-form frequency effects for compounds.

None of the other lexical predictors had a significant effect on first-and-only fixation durations. However, a factor that we have not taken into account thus far is contextual information. In the ET corpus compounds are presented in natural discourse contexts. These contexts potentially contain valuable information that helps predict and identify compounds. Unfortunately, quantifying the contribution of contextual information to compound recognition is not straightforward for the current data set. Compared to simple nouns, noun-noun compounds have a relatively low frequency of occurrence. As a result, it is hard to obtain reliable contextual predictability measures for the compounds in the ET corpus. The Google 1T n-gram data (Brants & Franz, 2006) contain all word trigrams with a frequency greater than or equal to 40 in a trillion word corpus. Nonetheless, the Google 1T trigram frequency list contains only 59.26% of all compound-final word trigram tokens in our data set. Similarly, the Google 1T trigram frequency list contains no more than 32.85% of all compound-initial trigram tokens in our data and 44.05% of all trigram tokens in which the compound is the middle word.

The fact that such a substantial proportion of the word trigrams tokens in our data set do not reach the frequency threshold of 40 per trillion in the Google 1T n-gram corpus suggests that trigram frequency may provide little explanatory power for a majority of the compound tokens in the ET corpus. Even if the average language user was to experience a billion words in her or his life, the binomial probability of ever encountering a word trigram with a frequency of 40 in the Google 1T n-gram corpus would be a mere 3.92% $(1 - P(X = 0|n = 1,000,000,000, p = \frac{40}{1,000,000,000,000}))$ and the chance of encountering such a word trigram more than once in a lifetime is no more than 0.08% $(1 - P(X = 0|n = 1,000,000,000, p = \frac{40}{1,000,000,000,000}) - P(X = 1|n = 1,000,000,000, p = \frac{40}{1,000,000,000,000}))$. For a majority of the compounds in the ET corpus data set it is there-

fore hard to see how contextual predictivity as gauged through trigram frequency measures could provide explanatory power for eye fixation patterns. To prevent a loss of statistical power due to missing data points we therefore decided to not include trigram frequency in our set of lexical predictors.

Nonetheless, contextual predictability may help compound recognition for the subset of compound trigrams that individual language users may have experienced. For the subset of word trigram tokens that appeared with a frequency greater than or equal to 40 in the Google 1T n-gram data, we therefore carried out a post-hoc analysis in which we included (log-transformed) trigram frequency in the Google 1T n-gram corpus as a fixed effect smooth in the lexical predictor GAMM reported above. This post-hoc analysis revealed a significant facilitatory effect of trigram frequency ($F = 7.453$, $p = 0.006$) on first-and-only fixation durations that was linear in nature.[1]

The effect of trigram frequency on first-and-only fixation durations confirms that contextual predictability co-determines eye fixation patterns during compound reading. The trigram frequency effect did not interact with either compound frequency or the LSA similarity between the compound and its head. The tensor product between Compound Frequency and LSA Similarity Compound-Head did not reach significance in this post-hoc analysis. A similar model on the full data set in which we set trigram frequency to 0 for compound-final trigrams that had a frequency smaller than 40 in the Google 1T n-gram data as well as for the 458 (10.10%) compound tokens that did not appear in compound-final trigrams (e.g., compounds that appear in the first two words of a sentence), however, showed a main effect of trigram frequency ($F = 21.764$, $p < 0.001$), as well as a significant tensor product interaction between Compound Frequency and LSA Similarity Compound-Head that was

[1] Similar post-hoc models on the subset of the data for which the compound-final trigram frequency was equal to or greater than 40 in the Google 1T n-gram data were fitted for all analyses reported below. In addition to the trigram frequency effect reported here, we found a trigram frequency effect for probability of refixation, as well as for first-of-many fixation position. We return to these effects when discussing the relevant analyses. No other significant effects of trigram frequency were observed.

quantitatively weaker ($F = 2.227$, $p = 0.041$), but qualitatively highly similar to the tensor product reported above.

3.4.1.2 NDL model. To find out how well Naive Discrimination Learning (NDL) is able to predict first-and-only fixation durations, we fit a similar GAMM with random effect terms for Participant ($F = 11.604, p < 0.001$) and Word ($F = 0.112, p = 0.040$) to the first-and-only fixation data. The model for the lexical predictors had shown main effects of Line, X Page and X Word, as well as an interaction between Compound Frequency and LSA Similarity Head-Compound. In the model based on the NDL measures, we again found significant main effects of Line ($F = 10.448, p < 0.001$), X Page ($F = 15.375, p < 0.001$) and X Word ($F = 2.860, p = 0.046$). All these effects were highly similar to the effects in the model with the lexical predictors. We therefore do not describe these effects in more detail here.

In the NDL model, the interaction between Compound Frequency and LSA Similarity Head-Compound is replaced by a simple main effect of NDL Activation Compound ($F = 4.485, p = 0.001$). Importantly, NDL Activation Compound is the integrative compound activation measure, which consists of a weighted sum of activation of the compound as a whole, the modifier and the head given the orthographic features of the compound. The fact that this integrative measure provides maximum explanatory power for the first-and-only fixation durations indicates that successful compound processing involves a rapid integration of all lexico-semantic information associated with the orthographic form of the compound. This lexico-semantic information is not limited to the compound lexeme, but includes the lexemes of the modifier and the head.

The main effect of NDL Activation Compound is presented in Figure 3.3. The effect of the NDL activation of the compound is inverse U-shaped in nature, with shorter fixation durations for compounds with either high or low activations in the NDL model. As such, the NDL activation of the compound shows an effect that is qualitatively similar to that of Compound Frequency.

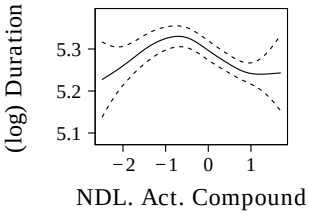

Figure 3.3. Effect of the NDL Activation Compound on (log) Fixation Duration of first-and-only fixations.

Neither the main effect of Compound Frequency ($F = 1.294$, $p = 0.263$), nor the main effect of LSA Similarity Head-Compound ($F = 1.600$, $p = 0.156$) reaches significance when the NDL activation of the compound ($F = 2.608$, $p = 0.032$) is added to the model. Furthermore, NDL Activation Compound interacts with neither Compound Frequency, nor LSA Similarity Head-Compound. This suggests that the effect of the LSA similarity between the compound and its head may not be purely semantic in nature, but instead (partly) reflect certain aspects of the bottom-up discriminability of a compound given its orthographic features.

Both models have similar REML scores (REML score lexical predictor model: 382.70, REML score NDL model: 383.33). The NDL model, however, uses 3 fewer degrees of freedom as compared to the lexical predictor model. As such, the current analysis indicates that the NDL framework provides a simpler, yet equally powerful account of the first fixation durations as do standard lexical predictors.[2]

When the interaction between Compound Frequency and LSA Similarity Head-Compound and the NDL activation of the compound are added to the model simultaneously, both effects remain significant (Compound Frequency by LSA Similarity Head-Compound: $F = 2.263$, $p = 0.038$; NDL Activation Compound: $F = 2.571$, $p = 0.036$). The REML score of this composite model (383.32), however, is similar to that of the

[2] Note, however, that the NDL model uses 1 weight parameter to define the activation of the compound. As such, it could be argued that the NDL model uses only 2 fewer degrees of freedom as compared to the lexical predictor model.

individual lexical predictor and NDL GAMMs. As such, adding both lexical predictor and NDL measures to the model does not provide additional explanatory value.

Given the significant effect of trigram frequency for the lexical predictor model, we carried out a post-hoc analysis in which we included the frequency of the compound-final word trigram as a predictor in the NDL model reported above. As before, we observed a significant effect of trigram frequency ($F = 9.494$, $p = 0.002$) for the subset of the data for which word-final trigram frequencies were available in the Google 1T n-gram corpus. This effect was linear in nature and did not interact with the effect of NDL Activation Compound. The main effect of NDL Activation Compound did not reach significance in this post-hoc analysis. A similar model on a larger data set with trigram frequency set to 0 for compound-final trigrams that had a frequency lower than 40 in the Google 1T n-gram data as well as for compound tokens that did not appear in compound-final trigrams, however, showed a significant main effect of NDL Activation Compound ($F = 2.888$, $p = 0.021$) in the presence of a trigram frequency effect ($F = 23.092$, $p <0.001$). This effect of NDL Activation Compound was qualitatively highly similar to the effect of NDL Activation Compound reported above.

Much like NDL measures of the activation of the compound provide a systemic alternative to lexical predictors that describe lexical properties of the compound, NDL measures of the contextual activation of the compound given the preceding two words could provide a systemic alternative to the trigram frequency measure. For the first-and-only fixation duration data reported here, however, NDL measures of the contextual predictability of the compound given the preceding two words did not provide explanatory power over and above the effect of NDL activation compound. Nonetheless, it is important to note that the architecture of the NDL model is fully consistent with n-gram frequency effects.

The orthography-to-lexeme network was trained with the orthographic features of the target word, as well as the orthographic features of the preceding two words as cues for the compound lexeme to allow for contextual learning. A number of previous studies have demonstrated that

this type of contextual learning allows discrimination learning networks to replicate n-gram frequency effects (see, e.g., Baayen et al., 2011, 2013). Furthermore, the lexeme-to-lexeme network allows for contextual learning effects at the lexico-semantic level. The absence of contextual learning effects in the NDL models reported here is likely to result from the limited size of the training lexicon for the NDL networks. At 100 million words, the British National Corpus may simply be too small to allow the NDL networks to learn the predictability of noun-noun compounds in context.

3.4.2 Single fixation position

The previous section described which lexical predictors and NDL measures had an influence on *how long* participants fixated on a compound during first-and-only fixations. Here, we investigate the influence of both types of predictors on *where* participants fixate.

We fitted a GAMM with random effects for Participant ($F = 46.857, p < 0.001$) and Word ($F = 0.128, p = 0.021$) to the first-and-only fixation positions. Similar to the model for first-and-only fixation durations, we observed main effects of Line ($F = 5.223, p < 0.001$) and X Page ($F = 14.701, p < 0.001$).

As expected, the effects of Line and X Page are in the opposite direction of the effects reported for fixation duration (see Figure 3.4). Participants move less far into the compound when the vertical position of the compound on the page is lower (i.e., the line number is greater) and move further into the compound when fixations are further to the right. The effect of X Page is particularly prominent near the right edge of the page, where a sharp increase in fixation position is seen. As for the effect of X Page on first-and-only fixation duration, this suggests that additional context on the same line helps read compounds more efficiently. The effects of the experimental predictors Line and X Page indicate that the preferred viewing position (Rayner, 1979) for first-and-only fixations is co-determined to a considerable extent by the physical position of the compound on a page.

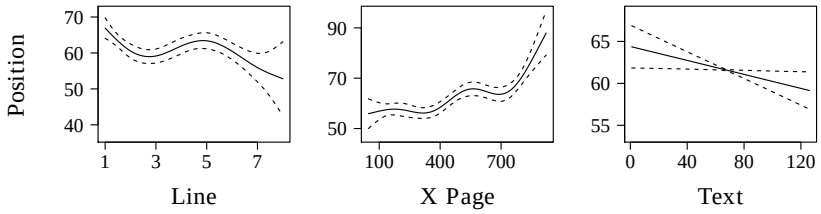

Figure 3.4. Effects of Line (left panel), X Page (middle panel) and Text (right panel) on Fixation Position of first-and-only fixations.

In addition to the effects of Line and X Page, we found a main effect of Text ($F = 4.837, p = 0.028$). As can be seen in the right panel of Figure 3.4, first fixation positions initially are relatively far into the compound. As the experiment proceeds, however, participants adapt their reading strategy and fixate less far into compounds.

Finally, we observed a significant effect of Compound Length ($F = 61.658, p < 0.001$) on first-and-only fixation positions. The effect of Compound Length is presented in Figure 3.5, which shows that the longer a compound, the more rightward the position of the fixation on that compound. This suggests that parafoveal preview (see, e.g., Rayner et al., 1982) allows participants to gauge the length of a compound before it is fixated on for the first time and to adjust the initial fixation position on the basis of this information. As such, the effect of Compound Length on first-and-only fixation position observed here fits well with the fact

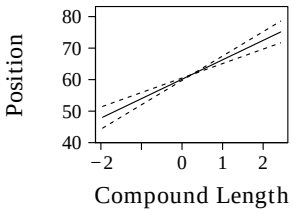

Figure 3.5. Effect of Compound Length on Fixation Position of first-and-only fixations.

that information about the length of a word acquired through parafoveal preview co-determines incoming saccade lengths (see, e.g., Rayner, 1979). We did not observe any effects of other lexical predictors or NDL measures on first-and-only fixation positions.

3.4.3 Probability of refixation

Thus far we investigated what happens when participants are able to process a compound in a single fixation. In the remainder of the Results section we take a look at fixation patterns when additional fixations are necessary. The current section forms a bridge between the single fixation and multiple fixation analyses and evaluates *when* additional fixations are necessary.

3.4.3.1 Lexical predictor model. We fitted a binomial GAMM with random effects for Participant ($\chi^2 = 19.401, p < 0.001$) and Word ($\chi^2 = 79.791, p = 0.003$) to the data. As can be seen in the left panel of Figure 3.6, we observed a main effect of X Page ($\chi^2 = 38.873, p < 0.001$) that was qualitatively similar to that observed for first-and-only fixation durations. As before, this effect of X Page suggests that additional preceding context on the same line allows for more efficient processing.

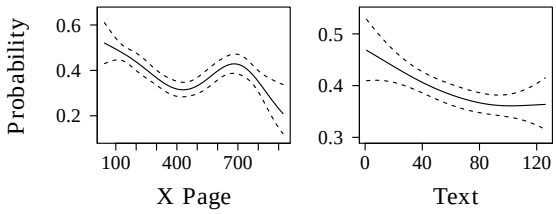

Figure 3.6. Effects of X Page (left panel) and Text (right panel) on Probability of Refixation.

In addition, we observed a main effect of Text ($\chi^2 = 9.975, p = 0.009$), with fewer refixation as participants move through the experiment (see right panel of Figure 3.6). This suggests that the leftward adjustment of the initial fixation position later in the experiment (see the analysis

of first-and-only fixation position above) helps process compounds more efficiently.

Furthermore, we observed effects of two lexical predictors. First, we observed a significant main effect of Compound Frequency ($\chi^2 = 20.074, p < 0.001$). This effect of Compound Frequency is presented in the left panel of Figure 3.7. Although confidence intervals are wide near the edges, the effect of compound frequency is facilitatory in nature, with a lower probability of refixation for high frequency compounds.

Second, we found an interaction of Compound Length with X Word ($\chi^2 = 197.928, p < 0.001$). This interaction is presented in the right panel of Figure 3.7. For fixations near the left border of a compound the probability of a refixation depends on the length of the compound: the greater the length of the compound, the greater the probability of a refixation. Interestingly, the probability of a refixation is also somewhat greater when participants fixate far into short compounds. The interaction of X Word and Length presented in Figure 3.7 is therefore characterized by a U-shaped curve along the main diagonal, in which the optimal fixation position within a compound is a function of the length of that compound: the greater the length of the compound, the more rightward the optimal viewing position (see, e.g., O'Regan, 1992; O'Regan & Jacobs, 1992, for optimal viewing position effects in isolated word reading)

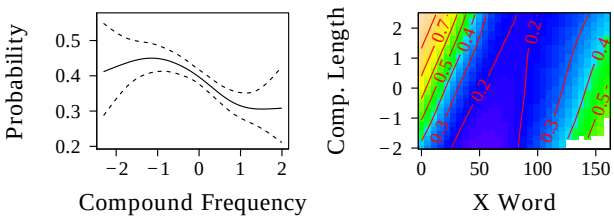

Figure 3.7. Effects of Compound Frequency (left panel) and the interaction between Compound Length and X Word (right panel) on Probability of Refixation.

A post-hoc analysis for the subset of the data with a compound-final trigram frequency of 40 or greater in the Google 1T n-gram data furthermore revealed a significant main effect of trigram frequency (χ^2 = 4.398, p = 0.043). This effect is near-linear in nature, with a greater probability of refixation for compounds that appear at the end of low-frequency trigrams. Contextual predictability therefore co-determines not only the duration of first fixations, but also the probability of an additional fixation. The effect of trigram frequency did not interact with Length or Compound Frequency.

While the effect of trigram frequency left intact the interaction of Length and X Word (χ^2 = 112.454, $p < 0.001$), it rendered the effect of compound frequency insignificant for the subset of the data for which trigram frequencies were available (χ^2 = 7.798, p = 0.097). A model on the full data set with a trigram frequency set to 0 when compound-final trigram frequencies were not available, however, showed a significant main effect of Compound Frequency (χ^2 = 13.940, p = 0.005) in the presence of a trigram frequency effect (χ^2 = 10.210, p = 0.001). This effect of Compound Frequency was qualitatively highly similar to the effect of Compound Frequency reported above. As for first-and-only fixation durations, therefore, the effect of trigram frequency on the probability of a second fixation seems to exist relatively independently of the effects of word-level lexical predictors.

3.4.3.2 NDL model. The lexical predictor model was characterized by a main effect of Compound Frequency and an interaction of Compound Length with X Word. To compare the performance of the NDL measures to that of the lexical predictors, we fitted a similar binomial GAMM with random effects for Participant ($\chi^2 = 19.267, p < 0.001$) and Word ($\chi^2 = 85.051, p = 0.001$) to the data. The fixed effects for X Page ($\chi^2 = 39.355, p < 0.001$) and Text ($\chi^2 = 9.826, p = 0.009$), as well as the interaction between X Word and Length ($\chi^2 = 197.465, p < 0.001$) remained significant in the NDL model and were qualitatively highly similar to the effects reported for the lexical predictor model.

The lexical predictor model had furthermore shown a main effect of Compound Frequency. In the NDL model this effect is replaced by an interaction between NDL Activation Compound and NDL MAD Compound ($\chi^2 = 21.512, p < 0.001$).[3] NDL MAD Compound quantifies the dispersion of the vector of compound weights given all word types in the lexicon. Conceptually, it can be thought of as a measure of the predictability of a compound that is independent of the current input. As such, the tensor product of NDL Activation Compound and NDL MAD Compound reflects an interaction between the bottom-up support for a compound and top-down knowledge about its prior probability.

The NDL model and the lexical predictor model have comparable UBRE scores (UBRE score lexical predictor model: 0.2066, UBRE score NDL model: 0.2064). One the one hand, this suggests that lexical predictors provide a simpler account of the data. On the other hand, upon closer inspection the tensor product interaction between NDL Activation Compound and NDL MAD Compound turns out to be a somewhat more precise measure of lexical processing as compared to Compound Frequency. When Compound Frequency and the interaction between NDL Activation Compound and NDL MAD Compound are entered into the model simultaneously, the tensor product remains significant ($\chi^2 = 18.410$, $p = 0.040$), whereas the effect of Compound Frequency loses significance ($\chi^2 = 8.397$, $p = 0.106$).

The NDL model furthermore provides information that the lexical predictor model does not. The interaction between NDL Activation Compound and NDL MAD Compound is presented in Figure 3.8. The probability of a refixation is highest when the NDL activation of the compound is high, but NDL MAD Compound is low – or, from an NDL perspective, when the orthography provides a lot of bottom-up support for a compound, but the prior probability of that compound is low. By contrast, the probability of a refixation is lowest when NDL Activation Compound and NDL MAD Compound are both high or – to a lesser extent – when both measures are low. This is expected when both NDL Activation Compound and NDL MAD Compound are high, which represents the ideal

[3] The interaction between NDL Activation Compound and NDL MAD Compound was restricted to 4 by 5 knots to prevent overfitting for extreme predictor values.

case for optimal processing. Perhaps surprisingly, however, the probability of a refixation also decreases when both NDL Activation Compound and NDL MAD Compound are low. This suggests that limited bottom-up support for a compound is unproblematic, as long as its prior probability is low as well, and vice versa. In other words: the performance of the language processing system is optimal when the information provided by the orthographic input is consistent with prior expectations based on the network connectivity of a compound.

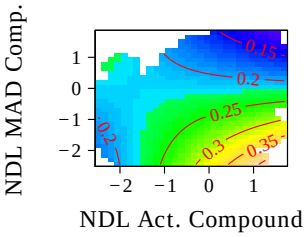

Figure 3.8. Effect of the interaction between NDL Activation Compound and NDL MAD Compound on Probability of Refixation.

Given the effect of trigram frequency for the lexical predictor model, we carried out a similar post-hoc analysis for the NDL model. In contrast to the lexical predictor model, however, we observed a marginally significant effect of trigram frequency only ($\chi^2 = 4.061$, $p = 0.057$). Furthermore, trigram frequency did not interact with NDL Activation Compound or NDL MAD Compound.

3.4.4 First-of-many fixation duration

While a majority of the compounds in our experiment were fixated on only once, 38.69% of all compounds required one or more additional fixations. In the next section we take a closer look at these additional fixations. In this section we first investigate what characterizes first-of-many fixations, by looking at their duration and position.

3.4.4.1 Lexical predictor model. To investigate the duration of first-of-many fixations we fitted a GAMM with a random effect for Participant ($F = 12.469, p < 0.001$) to the data. The random effect of Word was not significant and was therefore omitted from the model. As in the lexical predictor model for the first-and-only fixation duration data, we found significant main effects of Line ($F = 9.531, p = 0.002$), X Page ($F = 5.944, p = 0.002$)[4] and X Word ($F = 13.166, p < 0.001$). As can be seen in Figure 3.9, the effects of Line and X Word are in the same direction as those observed for first-and-only fixation durations. By contrast, while greater values of X Page resulted in shorter first-and-only fixation durations, the effect is reversed for first-of-many fixation durations – with longer fixation durations for greater values of X Page. This effect, however, is no longer significant when incoming saccade length is added to the model ($F = 1.941$, $p = 0.164$).

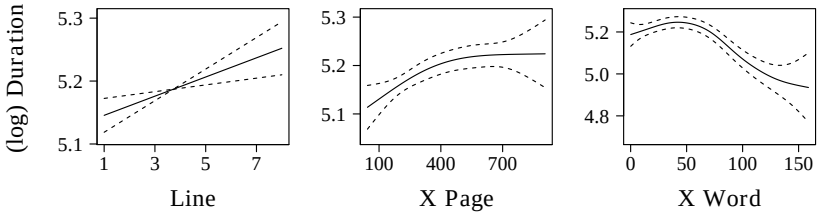

Figure 3.9. Effects of Line (left panel), X Page (middle panel) and X Word (right panel) on (log) Fixation Duration of first-of-many fixations.

Fixation patterns for first-and-only fixations were primarily determined by properties of the compound as a whole, such as Compound Frequency and Compound Length. The lexical predictors that influence first-of-many fixation durations, by contrast, are properties of the modifier. In particular, we found significant main effects of Modifier Length ($F = 11.375, p < 0.001$) and Modifier Frequency ($F = 6.133, p = 0.013$). As can be seen in the left panel of Figure 3.10, the effect of Modifier

[4] Even at 6 knots, the effect of X Page was characterized by an uninterpretable sinusoid pattern. We therefore limited the number of knots for the main effect smooth of X Page to 4.

Length is linear in nature, with longer fixation durations for compounds with longer modifiers. By contrast, the right panel of Figure 3.10 shows that higher values of Modifier Frequency lead to shorter fixation durations. This effect of Modifier Frequency, however, is only marginally significant when the magnitude of the incoming saccade is added to the model ($F = 2.811$, $p = 0.089$). As such, it is not clear how robust the effect of Modifier Frequency on first fixation durations is.

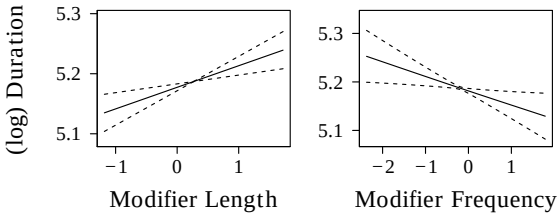

Figure 3.10. Effects of Modifier Length (left panel) and Modifier Frequency (right panel) on (log) Fixation Duration of first-of-many fixations.

The fact that lexical properties of the modifier rather than those of the compound as a whole co-determine first-of-many fixation durations gives a clear indication about the processing problems that lead to additional fixations. The results for first-and-single fixations suggest that optimal processing is comprehensive in nature. By contrast, the effects observed for first-of-many fixation durations are indicative of incomplete processing on the basis of partial information.

In the analysis for the probability of a second fixation we saw that lexical properties of the compound are one cause of incomplete processing: compounds that require only one fixation are shorter than those that require multiple fixations (8.54 characters versus 8.73 characters, $t = 3.823, p < 0.001$) and have higher frequencies (raw BNC frequency: 1022.58 versus 872.12, $t = -6.096, p < 0.001$).[5]

[5] The raw frequencies were log-transformed prior to statistical testing to remove a rightward skew from their distributions.

A second reason for suboptimal processing becomes clear from a closer inspection of the fixation positions. First-and-only fixations are further into the word than first-of-many fixations (59.94 pixels versus 50.79 pixels, $t = -6.994, p < 0.001$). This difference is reflected in smaller average forward saccade sizes for first-and-only fixations as compared to first-of-many fixations (143.26 pixels versus 150.61 pixels, $t = 2.923, p = 0.004$). As such, everything points towards an initial landing position that is too early in the compound, as a consequence of which the information necessary for efficient compound processing is not available.

The average fixation duration for first-of-many fixations is shorter than that for first-and-only fixations (175.16 ms versus 195.13 ms, $t = -8.477, p < 0.001$). This indicates that participants realize that successful processing will not be possible during the current fixation and rapidly move on in an attempt to complement the incomplete information that becomes available during first-of-many fixations with additional information that is obtained through a second fixation.

3.4.4.2 NDL model. To investigate the explanatory power of the NDL measures for first-of-many fixation durations, we fitted a GAMM with a random effect for Participant ($F = 12.370, p < 0.001$) to the data. As in the lexical predictor model, we observed significant main effects of Line ($F = 9.743, p = 0.002$), X Page ($F = 5.952, p = 0.002$) and X Word ($F = 13.313, p < 0.001$). As for the lexical predictor model, too, the effect of X Page is no longer significant when incoming saccade length is added to the model ($F = 1.267$, $p = 0.240$). The effect of Modifier Length also remained significant in the NDL model ($F = 10.064, p = 0.002$) and was qualitatively highly similar to the effect of Modifier Length in the lexical predictor model.

More interestingly, the effect of Modifier Frequency that was present in the lexical predictor model is replaced by an effect of NDL Self-Activation Modifier ($F = 4.642, p = 0.031$) in the NDL model. Similar to the effect of Modifier Frequency, the effect of NDL Self-Activation Modifier is facilitatory in nature, with shorter first-of-many fixation durations when the modifier has a higher activation (see Figure 3.11). Importantly, it

is the NDL activation of the modifier given the orthographic features of the modifier, rather than the activation of the modifier given the orthographic features of the compound as a whole that proved most predictive. This fits well with the fact that fixation positions were more leftward for first-of-many fixations as compared to first-and-only fixations and suggests that the orthographic features of the head were not yet available during first-of-many fixations on compounds.

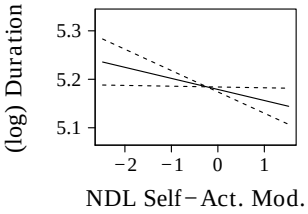

Figure 3.11. Effect of NDL Self-Activation Modifier on (log) Fixation Duration of first-of-many fixations.

A post-hoc analysis indicated that not all features of the modifier may be available either during first-of-many fixations. For this post-hoc analysis we calculated the NDL activation of the modifier lexeme given the first trigram of the compound only (i.e., *#ha* for the compound *handbag*). This measure turned out to be a highly significant predictor of first-and-only fixation duration ($F = 10.190, p = 0.001$), showing a linear effect that was qualitatively highly similar to that reported for NDL Self-Activation Modifier above. The activation of the modifier lexeme given the first orthographic trigram had increased explanatory power over NDL Self-Activation Modifier. When both measures were entered into the model simultaneously, the effect of NDL Self-Activation Modifier was no longer significant ($F = 1.572$, $p = 0.210$), whereas the activation of the modifier given the first orthographic trigram remained highly significant ($F = 7.371$, $p = 0.007$).

Furthermore, whereas the effect of NDL Self-Activation Modifier was only marginally significant when incoming saccade length was added to the model as a predictor ($F = 3.467$, $p = 0.063$), the significance of the

activation of the modifier lexeme given the first orthographic trigram was unaffected by incoming saccade length ($F = 7.850$, $p = 0.005$). Similarly, whereas the effect of modifier length remained marginally significant when incoming saccade length was added to the original NDL model ($F = 3.611$, $p = 0.058$), it was no longer significant when incoming saccade length was added to the post-hoc NDL model described here ($F = 2.589$, $p = 0.108$). As such, the activation of the modifier lexeme given the first orthographic trigram is a superior predictor of first-and-only fixation duration as compared to NDL Self-Activation Compound. This suggests that on average only a very limited subset of the orthographic features of the compound is available during first-of-many fixations.

When comparing the lexical predictor model with the original NDL model including NDL Self-Activation Modifier, both models show similar explanatory power, with comparable REML scores for the lexical predictor model (325.92) and the NDL model (326.65). When both Modifier Frequency and NDL Self-Activation Modifier were added to the model simultaneously, neither Modifier Frequency ($F = 1.900$, $p = 0.178$), nor NDL Self-Activation Modifier ($F = 0.292$, $p = 0.589$) reached significance. The marginal differences in performance between both models are unsurprising given the fact that the Modifier Frequency and NDL Self-Activation Modifier measures are highly correlated ($r = 0.706$).

When comparing the lexical predictor model to the NDL model that includes the activation of the modifier given the first orthographic trigram, rather than NDL Self-Activation Modifier, however, the REML score of the NDL model is somewhat better than that of the lexical model (lexical model: 314.75[6]; NDL model: 312.31). In addition, whereas the effect of the NDL activation of the modifier given the first orthographic trigram remained significant ($F = 6.496$, $p = 0.011$) when both predictors were entered into a model simultaneously, the effect of Modifier Frequency was no longer significant ($F = 1.216$, $p = 0.270$). As such, the NDL framework

[6] Note that this REML score is different from the one reported above. This is a consequence of simultaneously removing outliers for the NDL and lexical predictor models. The original lexical predictor model did not include NDL Self-Activation outliers, whereas the lexical predictor model reported here did not include outliers for the activation of the modifier given the first orthographic trigram.

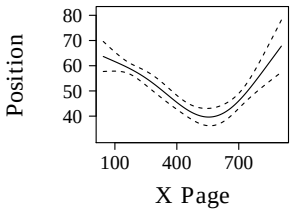

Figure 3.12. Effect of X Page on Fixation Position of first-of-many fixations.

provides a competitive theory for understanding the first-of-many fixation duration data, particularly when the limited amount of orthographic information that is available during these fixations is taken into account.

3.4.5 First-of-many fixation position

A GAMM with random effects for Participant ($F = 28.522, p < 0.001$) and Word ($F = 0.245, p < 0.001$), showed a significant effect of X Page ($F = 18.693, p < 0.001$).[7] This effect of X Page is presented in Figure 3.12. For first-and-only fixations, we saw an increase in first fixation position when the horizontal position of a compound on a page was greater. For first-of-many fixations, the effect is U-shaped in nature.

Similar to the effect of X Page for first-and-only fixation position, participants fixate further into compounds near the right edge of the page, presumably due to the increased amount of contextual information on the same line. In addition, however, first-of-many fixation positions are also further into compounds that appear near the left edge of the page. Whereas sub-optimal initial fixation positions typically are *not far enough* into the compound, the U-shaped effect of X Page observed here suggests that sub-optimal initial fixation positions on compounds that are at or near the beginning of a new line tend to be *too far* into the compound.

[7] As for the effect of X Page on first-of-many fixation duration, even at 6 knots the effect of X Page on first-of-many fixation position was characterized by an uninterpretable sinusoid pattern. As before, we therefore limited the number of knots for the main effect smooth of X Page to 4.

The fixation position data for first-and-only fixations showed a significant effect of the length of the compound. No such length effect was observed for first-of-many fixation position. Furthermore, no other lexical predictors or NDL measures were predictive for the fixation position of first-of-many fixations. The absence of significant effects for fixation position of first-of-many fixations fits well with the increased average saccade size for first-of-many fixations as compared to first-and-only fixations (150.61 pixels versus 143.26 pixels). While parafoveal preview enabled participants to commence compound processing prior to first-and-only fixations, the increased distance from the previous fixation to the compound substantially reduced the extent to which parafoveal pre-processing was possible prior to first-of-many fixations.

While we found no evidence for effects of lexical predictors or NDL measures, we did find some support for an effect of trigram frequency. Whereas the effect of trigram frequency was marginally significant only for first-and-only fixation position ($F = 3.547$, $p = 0.060$), a post-hoc analysis on the subset of the data for which trigram frequencies were available in the Google 1T n-gram data showed a significant linear main effect of trigram frequency ($F = 3.900$, $p = 0.049$) on first-of-many fixation position, with more *leftward* fixation positions for compounds that appeared in more frequent compound-final trigrams. A similar model on the full data set, in which we set trigram frequency to 0 for compounds for which trigram frequencies were not available, however, did not show a significant effect of trigram frequency ($F = 1.604$, $p = 0.234$). The effect of trigram frequency on first-of-many fixation position, therefore, is not statistically robust.

3.4.6 Second fixation duration

The analysis of first-of-many fixations indicated that additional fixations are necessary when processing during the initial fixation has to proceed on the basis of partial information. In this section, we investigate the fixation durations and fixation positions of second fixations to gauge how the situation of having insufficient information during the initial fixation is corrected for.

3.4.6.1 Lexical predictor model. In addition to a random effect for Participant ($F = 8.318, p < 0.001$), the lexical predictor GAMM showed main effects of Line ($F = 13.843, p < 0.001$) and X Page ($F = 7.786, p < 0.001$). The main effects of Line and X Page are presented in Figure 3.13 and are similar in nature to the effects of Line and X Page for first-of-many fixation durations. As for first-of-many fixation durations, fixations near the bottom of a page were characterized by longer fixation durations. The effect of X Page is quadratic in nature and shows shorter fixation durations for second fixations near the right of the page. As before, this suggests that additional sentential context reduces fixation durations.

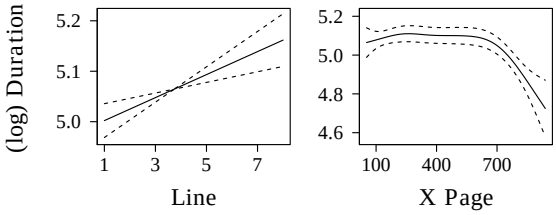

Figure 3.13. Effects of Line (left panel) and X Page (right panel) on (log) Fixation Duration of second fixations.

The lexical predictor GAMM furthermore showed a significant main effect of Compound Frequency ($F = 7.848, p = 0.005$). This effect of Compound Frequency is presented in Figure 3.14, which shows that second fixation durations are shorter for high frequency compounds than for their low frequency counterparts.

While first-of-many fixation durations were determined by lexical properties of the modifier, the only lexical predictor that proved significant for second fixation durations is a lexical property of the compound itself. Similar to processing during first-and-only fixations, therefore, processing during second fixations is driven by lexico-semantic information associated with the compound as a whole. Second fixations, then, are perhaps best conceived as a renewed attempt to process a compound in an optimal, comprehensive fashion. Given that part of the required information was already available through the first fixation, however, this

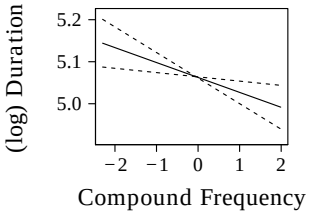

Figure 3.14. Effect of Compound Frequency on (log) Fixation Duration of second fixations.

second pass is completed substantially faster than processing during first-and-single fixations, as indicated by shorter average fixation durations (158.53 ms versus 195.13 ms, $t = -14.451, p < 0.001$).

3.4.6.2 NDL model. Similar to the lexical predictor model, the NDL model for second fixation duration showed a significant random effect for Participant ($F = 8.395, p < 0.001$) as well as significant fixed effect smooths for Line ($F = 13.637, p < 0.001$) and X Page ($F = 7.810, p < 0.001$). The effects of Line and X Page were qualitatively highly similar to the effects of Line and X Page in the lexical predictor model.

In the NDL model, the effect of Compound Frequency is replaced by a significant main effect of NDL MAD Compound ($F = 8.639, p = 0.003$). As can be seen in Figure 3.15, the effect of NDL MAD Compound is linear in nature and qualitatively highly similar to the effect of Compound Frequency. This similarity is reflected in the comparable explanatory power of the lexical predictor model and the NDL model. Both models had highly similar REML scores (REML score lexical predictor model: 529.85, REML score NDL model: 529.37) and when both predictors were added to the model simultaneously, neither Compound Frequency ($F = 0.877$, $p = 0.349$), nor NDL MAD Compound ($F = 1.655$, $p = 0.199$) reached significance. The comparable performance of both models is unsurprising given the fact that Compound Frequency and NDL MAD Compound are extremely highly correlated ($r = 0.993$) As for first-of-many fixation durations, therefore, the results for second fixation durations indicate

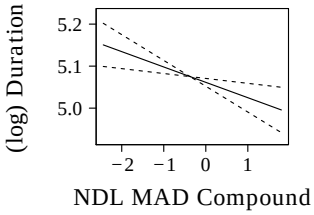

Figure 3.15. Effect of NDL MAD Compound on (log) Fixation Duration of second fixations.

that the quantitative performance of NDL measures and standard lexical predictors is similar when multiple fixations are required.

Despite the fact that the explanatory power of the lexical predictor model and the NDL model is similar for the second fixation duration data, the NDL model provides valuable information about the nature of compound processing during second fixations that is not available from the lexical predictor analysis. In the lexical predictor model we observed an effect of Compound Frequency for both first-and-only fixation duration and second fixation duration. The NDL model, however, demonstrates that the processing mechanisms that underlie both effects of Compound Frequency are markedly different. First-and-only fixation durations are co-determined by NDL Activation Compound. This suggests that during first-and-only fixations readers rely primarily on the bottom-up support for a compound and its constituents from the orthographic features in the visual input. For those compounds that require a second fixation, however, the bottom-up information that was available during the first fixation proved insufficient for successful compound processing. The fact that NDL MAD Compound co-determines second fixation durations indicates that rather than making a second attempt at processing the compound in a bottom-up fashion, readers fall back on top-down best guesses based on the prior probability of lexical candidates.

3.4.7 Second fixation position

Other than an effect of Compound Length for first-and-only fixations, the fixation position models for first-and-only and first-of-many fixation positions provided little evidence for effects of lexical predictors or NDL measures. In this section we investigate whether or not lexical predictors or NDL measures do co-determine second fixation positions.

3.4.7.1 Lexical predictor. A GAMM with random effects for Participant ($F = 12.236, p < 0.001$) and Word ($F = 0.200, p = 0.004$) again showed a main effect of X Page ($F = 28.467, p < 0.001$). This effect of X Page is presented in the left panel of Figure 3.16, which shows that fixations are further into the word for compounds that appear near the right edge of the page. Once more, this suggests that additional sentential context allows for more efficient processing.

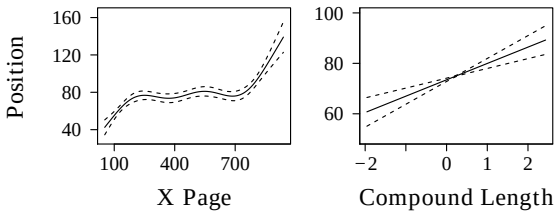

Figure 3.16. Effect of X Page (left panel) and Compound Length (right panel) on Fixation Position of second fixations.

In addition to the effect of X Page, we also observed an effect of Compound Length ($F = 24.991, p < 0.001$) As can be seen in the right panel of Figure 3.16, fixations are more towards the right for longer compounds. This fits well with the X Word by Compound Length interaction observed for probability of refixation. The optimal viewing position varies as a function of compound length: the longer a compound, the more rightward the optimal viewing position.

As for the effects on second fixation duration, the effect of Compound Length on second fixation position indicates that properties of the compound as a whole, rather than properties of (one of) the constituents

determine fixation patterns for second fixations. As such, the second fixation position data indicate that processing during second fixations is to a larger extent driven by the compound lexeme as compared to processing during first-of-many fixations. No other lexical predictors showed a significant effect on second fixation position.

3.4.7.2 NDL model. A GAMM with random effects for Participant ($F = 12.031, p < 0.001$) and Word ($F = 0.154, p = 0.018$), showed main effects of X Page ($F = 29.308, p < 0.001$) and Compound Length ($F = 28.423, p < 0.001$). These effects were highly similar to the effects of X Page and Compound Length reported for the lexical predictor model above.

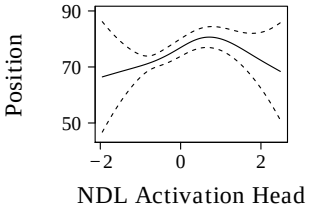

Figure 3.17. Effect of NDL Activation Head on Fixation Position of second fixations.

In addition to the effects of X Page and Compound Length, we found some evidence for an effect of NDL Activation Head (i.e., the activation of the lexeme of the head given the orthographic features of the compound). The effect of NDL Activation Head ($F = 3.138, p = 0.021$) is presented in Figure 3.17. Near the middle of the NDL Activation Head range, there is an upward trend, with second fixation position being further into the word when the activation of the head is high. At the edges of the predictor range, however, there is considerable uncertainty about the nature of the effect. Furthermore, when incoming saccade length was added to the model, the effect of NDL Activation Head no longer reached significance ($F = 2.271$, $p = 0.085$). It is, therefore, unclear how robust the effect of NDL Activation Head is.

3.4.8 Probability of third fixation

The previous sections demonstrated that processing during first-of-many fixations is based on partial bottom-up information and therefore suboptimal, and that the additional information that becomes available during second fixations allows for processing that is better optimized. Two fixations were sufficient to process a vast majority of all compounds. For 285 (9.14%) compound tokens, however, a third fixation proved necessary. In this section we investigate which lexical predictors and NDL measures co-determine the probability of a third fixation.

3.4.8.1 Lexical predictor model. A GAMM with random effects for Participant ($\chi^2 = 25.336, p < 0.001$) and Word ($\chi^2 = 48.576, p = 0.013$) revealed a significant main effect of X Word ($\chi^2 = 36.100, p < 0.001$), with a lower probability of a third fixation if second fixations were further into the compound (see Figure 3.18). This effect of X Word demonstrates again that incomplete information due to a fixation position that is too far towards the left leads to additional fixations.

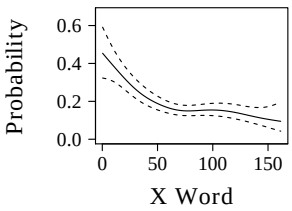

Figure 3.18. Effect of X Word on Probability of Third Fixation.

The lexical predictor GAMM furthermore showed a significant interaction between Compound Length and Compound Frequency ($\chi^2 = 22.198, p < 0.001$). The interaction between Compound Length and Compound Frequency is presented in the left panel of Figure 3.19, which shows that the probability of a third fixation is increased only for compounds that are both long and infrequent.

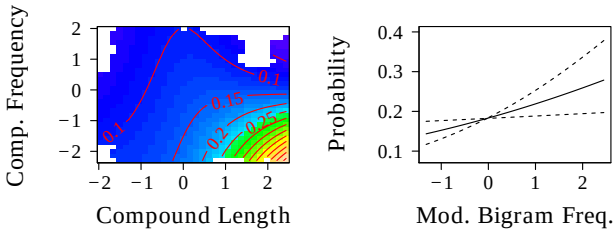

Figure 3.19. Effects of the interaction between Compound Length and Compound Frequency (left panel) and Modifier Mean Bigram Frequency (right panel) on Probability of Third Fixation.

Finally, we found some evidence for an effect of Modifier Mean Bigram Frequency ($\chi^2 = 5.863, p = 0.015$). As can be seen in the right panel of Figure 3.19, a higher average frequency of the orthographic bigrams in the modifier leads to more third fixations. When incoming saccade length was added to the model, however, the effect of Modifier Mean Bigram Frequency no longer reached significance ($\chi^2 = 8.672$, $p = 0.127$).

3.4.8.2 NDL model. As for the lexical predictor model, the NDL model showed significant random effects for Participant ($\chi^2 = 25.614, p < 0.001$) and Word ($\chi^2 = 47.622, p = 0.015$), as well as a main effect of X Word ($\chi^2 = 36.013, p < 0.001$). In addition, the NDL model showed a main effect of Modifier Mean Bigram Frequency that was significant in the base model ($\chi^2 = 5.882, p = 0.015$), but that lost significance when the incoming saccade size was added to the model ($\chi^2 = 9.090$, $p = 0.117$).

In the NDL model the interaction between Compound Length and Compound Frequency is replaced by an interaction between Compound Length and NDL MAD Compound ($\chi^2 = 21.803, p < 0.001$) As can be seen in Figure 3.20, the interaction between Compound Length and NDL MAD Compound is highly similar to the interaction between Modifier Mean Bigram Frequency and Compound Frequency, with a greater probability of a third fixation for long compounds with a low prior probability.

Analogous to the NDL model for second fixation duration, the NDL measure that replaced Compound Frequency was NDL MAD Compound, rather than NDL Activation Compound. As before, this suggests that while early fixation patterns are co-determined by the bottom-up support for a compound given the orthographic cues in the input, later measures are influenced by the prior probability of a compound.

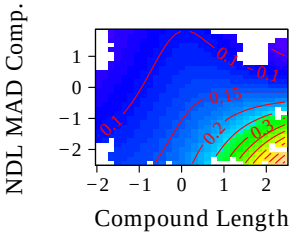

Figure 3.20. Effect of the interaction between Compound Length and NDL MAD Compound on Probability of Third Fixation.

The explanatory power of the NDL model was comparable to that of the lexical predictor model (UBRE score NDL model: -0.0594; UBRE score lexical predictor model: -0.0594). When Compound Frequency and NDL MAD Compound were entered into the model simultaneously, the effect of NDL MAD Compound remained significant ($\chi^2 = 3.877$, $p = 0.049$), whereas the effect of Compound Frequency lost significance ($\chi^2 = 2.945$, $p = 0.101$). Conversely, the interaction of Compound Length with Compound Frequency remained marginally significant ($\chi^2 = 7.239$, $p = 0.051$), whereas the interaction with Compound Length with NDL MAD Compound was no longer significant ($\chi^2 = 1.609$, $p = 0.205$).

3.4.9 Third fixation duration

In the previous section we investigated when third fixations are necessary. We saw that no more than 9.14% of compound tokens required a third fixation. The statistical power for the third fixation data therefore, is limited. Nonetheless, we decided to include the results for the third fixation data in this write-up. The reader is advised, however, that the

results reported here may not be as statistically robust as the effects reported for first and second fixations above.

A GAMM model on the third fixation duration data showed no significant random effects for Participant or Word. We did, however, observe a significant effect of Text ($F = 3.843, p = 0.014$). As can be seen in Figure 3.21, third fixation durations are shorter near the end of the experiment. This suggests that throughout the experiment participants learn to reduce the costs of having to fixate on compounds a third time.

Furthermore, we found a significant effect of LSA Similarity Modifier-Compound ($F = 5.472, p = 0.020$): the more semantically similar a modifier is to the compound as a whole, the shorter the duration of the third fixation (see right panel of Figure 3.21).

Although the robustness of the effect of LSA Similarity Modifier-Compound for the third fixation duration data is unclear given the limited number of third fixation in the current data, a post-hoc analysis on the next fixation data for single fixation cases (i.e., fixations following first-and-only fixations) provided independent support for late semantic effects.

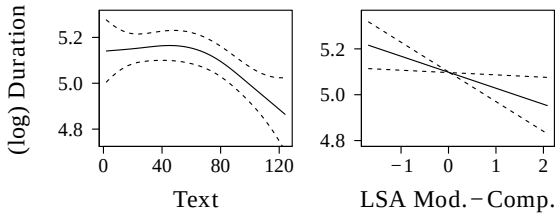

Figure 3.21. Effects of Text (left panel) and LSA Similarity Modifier-Compound (right panel) on (log) Fixation Duration of third fixations.

A GAMM model on the next fixation duration data in which we controlled for the length and frequency of the word that was fixated on showed an effect of LSA Similarity Modifier-Compound ($F = 16.844$, $p < 0.001$) that was qualitatively highly similar to the effect of LSA Similarity Modifier-Compound reported here. In addition, both the next fixation duration ($F = 4.488$, $p = 0.034$) and next fixation position ($F = 5.473$, $p = 0.019$) for first-and-only fixations showed significant effects

of LSA Similarity Modifier-Head. The current data therefore provide some support for late semantic effects regarding the similarity between the modifier on the one hand and the head and compound on the other hand.

We found no other effects of lexical predictors or NDL measures on third fixation duration. In the post-hoc analysis of the next fixation duration for first-and-only fixations, however, we found a linear effect of NDL Self-Activation Modifier (i.e., the activation of the modifier lexeme given the orthographic features of the modifier only), with shorter fixation durations for higher values of NDL Self-Activation Modifier ($F = 6.825$, $p = 0.009$). This effect of NDL Self-Activation Modifier provides some additional evidence for the idea that properties of the modifier co-determine later measures of compound processing.

3.4.10 Third fixation position

A GAMM with a random effect of Participant ($F = 2.546, p = 0.014$) showed significant effects of X Page ($F = 3.485, p = 0.004$) and Compound Length ($F = 14.468, p < 0.001$). As can be seen in Figure 3.22, the effects of X Page and Compound Length are extremely similar to the effects of both predictors for second fixation position. The processes that guide the fixation position of second and third fixations therefore seem to be highly similar. No other effects of lexical predictors or NDL measures on third fixation position were observed.

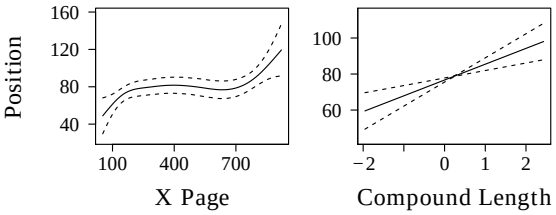

Figure 3.22. Effect of X Page (left panel) and Compound Length (right panel) on Fixation Position of third fixations.

3.5 General Discussion

We investigated eye fixation patterns for noun-noun compounds in the Brown corpus fiction section of the Edmonton-Tübingen eye-tracking corpus. Despite the fact that compounds are relatively infrequent in the English language, the over 920,000 fixations in this part of the ET corpus allowed us to look at compound processing in discourse context for 4747 fixations on 844 compound tokens. While compound processing has received considerable attention in the experimental psycholinguistic literature, to our knowledge the current work is the first to address compound processing in discourse context on this scale.

Following previous work, we investigated the eye fixation patterns for the ET corpus data using a wide range of lexical predictors, including frequency measures, neighborhood density, morphological family size and semantic similarity measures. Additionally, we investigated to what extent systemic measures, derived from Naive Discrimination Learning (henceforth NDL; Baayen et al., 2011) could help further understand the eye fixation data in the ET corpus. Previously, the NDL model has been applied primarily to unidimensional data obtained in behavioral experiments (see, e.g., Baayen et al., 2011, 2013; Ramscar et al., 2014). The current work is a first exploration of the explanatory power of NDL measures for eye-tracking data.

Table 3.1 provides a summary of the results of the present investigations. In what follows, we first discuss the results of the lexical predictor analysis, as well as the implications of this analysis for the different types of compound processing models that have been proposed in the literature. Next, we take a closer look at the results of the NDL analysis and the additional insights this analysis provides into the nature of compound processing.

The current findings indicate that presenting compounds in unnatural experimental tasks has a considerable influence on the nature of compound processing. In comparison to previous experimental work, compound processing is much more efficient in the ET corpus data. For all 4 participants in the subset of the ET corpus used here, a majority

Table 3.1. Summary of lexical predictor and NDL analyses. Numbers indicate p-values for the corresponding model terms. Round brackets show p-values when trigram frequency is added to the model; square brackets show p-values when incoming saccade length is added to the model.

	lexical	NDL
first-and-only fixation duration		
Comp. Freq. by LSA Head-Comp.	0.003 (0.041)	
NDL Activation Compound		0.001 (0.021)
first-and-only fixation position		
Compound Length	<0.001	<0.001
probability of second fixation		
Compound Frequency	<0.001 (0.005)	
Compound Length by X Word	<0.001	<0.001
NDL Act. Comp. by MAD Comp.		<0.001
first-of-many fixation duration		
Modifier Length	<0.001	0.002 [n.s.]
Modifier Frequency	0.013 [0.089]	
NDL Act. Mod. given first trigram		0.001
second fixation duration		
Compound Frequency	0.005	
NDL MAD Compound		0.003
second fixation position		
Compound Length	<0.001	<0.001
NDL Activation Head		0.021 [0.085]
probability of third fixation		
Mod. Mean Bigram Frequency	0.015 [n.s.]	0.015 [n.s.]
Comp. Length by Comp. Freq.	<0.001	
Comp. Length by NDL MAD Comp.		<0.001
third fixation duration		
LSA Modifier-Compound	0.020	
third fixation position		
Compound Length	<0.001	<0.001

of the compounds was processed using a single fixation (range: 54.06% - 66.71%, mean: 61.31%).

As can be seen in Table 3.1, fixations patterns for these first-and-only fixations were characterized by lexical properties and NDL measures of the compound as a whole, with early effects of compound length and compound frequency (cf. Zwitserlood, 1994; Bertram & Hyönä, 2003; Andrews et al., 2004; Kuperman et al., 2009). As we found out in a post-hoc analysis on the subset of the data for which trigram frequencies were available in the Google 1T *n*-gram data, there is a real possibility that word trigram frequency also predicts first and only fixation durations, in which case the evidence for any constituent involvement, coming from the LSA compound-head similarity measure in interaction with compound frequency, becomes much less convincing ($p = 0.041$).

The pattern of results for single fixation cases fits remarkably well with non-decompositional theories of lexical processing (Butterworth, 1983; Janssen et al., 2008) and theories in which whole-word access representations are claimed to play a role in lexical processing from the very beginning (see, e.g., Kuperman et al., 2009). The results for first-and-only fixations are also in line with supra-lexical models of compound processing, which propose that full-form access precedes access to the constituent morphemes (see, e.g., Giraudo & Grainger, 2001).

For words read with multiple fixations, modifier length and – to a lesser degree – modifier frequency predicted the initial fixation duration. By contrast, second fixation durations and positions were determined by the length and frequency of the compound as a whole. For third fixations, the fixation position is predicted again by compound length, whereas duration may have been influenced by the LSA modifier-compound similarity measure. Thus, the analysis based on lexical-distributional variables suggests that multiple-fixation trials are initiated with modifier-driven processing, followed by whole-word look-up, in turn followed by further semantic integration. This scenario for the nearly 40% of cases with multiple-fixation reading fits remarkably well with the stages in lexical access proposed by left-dominant sub-lexical models of compound processing, which hold that access to the modifier is essential for and

temporally precedes access to the full-form (Taft & Forster, 1975, 1976; Taft, 1979, 1991, 2004).

The combined evidence for single fixation trials and multiple fixation trials, however, does not fit very well with the above-mentioned theories of compound processing. Non-decompositional or supra-lexical models of compound processing do not fit well with the modifier-driven processing for multiple fixation trials. At best, therefore, non-decompositional or supra-lexical models describe a best-case scenario. By contrast, neither left-dominant, nor right-dominant (Juhasz et al., 2003; cf. Juhasz, 2007) sub-lexical models can straightforwardly account for the whole-form effects on first-and-only fixation duration and position.

The lexical predictor analysis also challenges dual route models. In parallel dual-route models a compound and its constituents are activated simultaneously. Typically, parallel dual-route models propose a horse race between the full-form route and the decompositional route, which operate in a simultaneous and independent fashion (see, e.g., Schreuder & Baayen, 1995; Baayen & Schreuder, 1999; Allen & Badecker, 2002; cf. Baayen & Schreuder, 2000, however, for a dual-route model that allows for an interaction of both processing routes). The pattern of results observed here provides little evidence for a horse race model, in which a parallel pursuit of a decompositional and full-form route is the default processing mechanism for compounds.

When compounds were processed in a single fixation, we did not observe simultaneous effects of constituents and full-form properties and found evidence for late semantic constituent effects only. For compounds that required multiple fixations, we observed effects of lexical properties of both the left constituent and the full form. The effects of left constituent properties and compound properties, however, were strictly separated in time, with left constituent effects characterizing first-of-many fixations and full-form effects characterizing second fixations. Parallel dual-route models, by contrast, would predict the effects of left constituent properties and full-form properties to temporally coincide (see, however, Bertram & Hyönä, 2003 for a dual-route model that allows for a head start of the decompositional route).

The only statistical model in which we observed simultaneous effects of constituent and full-form properties was that for the probability of a third fixation. The effect of the average bigram frequency of the modifier, however, was no longer significant when the size of the incoming saccade was taken into account. Even if the effect of average bigram frequency were statistically robust, however, we would, much rather interpret this late co-occurrence of left constituent and full-form effects as an attempt at a top-down integration of the incomplete first pass and the comprehensive second pass at compound processing than as evidence for a bottom-up dual-route architecture.

Libben (2006) argued that the architectures of existing models of morphological processing may be too restrictive (cf. Kuperman et al., 2009) and proposed a maximization of opportunity account of compound processing, in which readers simultaneously use all opportunities for compound processing that are available in the input. The lexical predictor analysis is not incompatible with Libben (2006). When all orthographic features are available in the input and the compound is not too long or infrequent, a single fixation suffices to successfully process the compound. When only a subset of the orthographic features is available during a first fixation due to suboptimal fixation planning, additional fixations are necessary. As indicated by the substantially shorter durations of second fixations as compared to first-and-only fixations (158.53 ms versus 195.13 ms), however, participants still obtain as much information as possible during first-of-many fixations and use this information to process compounds more efficiently during second fixations.

Within the theoretical idea of a maximization of opportunities model of compound processing, however, there is considerable room for different implementations of a model. One potential implementation of a maximization of opportunities model would be an interactive activation model in which the proposed sub-lexical, supra-lexical and dual route architectures of existing models are combined in a multiple route interactive activation model. In the NDL framework proposed here, however, we propose that compound processing is based on the information that becomes available through a single probabilistic learning mechanism.

The current NDL approach is based on two discrimination learning networks: an orthography-to-lexeme network and a lexeme-to-lexeme network. The orthography-to-lexeme network provides an estimate of the bottom-up support for a compound given the orthographic information in the visual input. The current pattern of results indicates that readers use all orthographic information that is available in the spotlight of visual attention to process compound words. The more complete the orthographic information for a compound, the better readers' chances are to correctly identify the compound during a single fixation.

First-and-only fixations were co-determined by an effect of the lexico-semantic information associated with a compound given the orthographic features of the compound as a whole. When there is sufficient bottom-up information to activate the lexemes of the compound and its constituents, therefore, no further fixations are required. Evaluation of how these lexemes contribute to the understanding of the discourse, however, is not completed by the end of the first-and-only fixation, as witnessed by a spill-over effect of the LSA similarity between the modifier and the compound on the fixation duration for the next word.

First-of-many fixations are characterized by more leftward fixation positions and longer incoming saccade lengths as compared to first-and-only fixations. This results in decreased parafoveal pre-processing of the compound and incomplete bottom-up information during initial fixations. For first-of-many fixations on a compound at the start of a new line, the pattern is reversed, with first-of-many fixation positions that are *too far* into the compound as a result of overshooting an optimal viewing position.

The suboptimal viewing position for first-of-many fixation durations results in insufficient bottom-up information for successful single fixation processing. The modifier-driven access suggested for the first-of-many fixation durations by the lexical-distributional measures, is replaced by a low-level effect in the NDL analysis: the weight on the connection from the compounds' initial trigram to the modifier lexeme. Since for multiple-fixation compound reading, compounds tend to be longer and to have lower frequencies, the initial fixation position is too early in the

word to allow the bottom-up information for the compound lexemes to be effective. All that happens at this stage is that the modifier lexeme receives support, limited to what its very first letter trigram can provide. Rather than proposing two separate strategies, one involving holistic processing and one involving decompositional processing, therefore, we propose that compound processing is determined by the way a single system responds to the input that it has at its disposal.

Clearly, the implementation of visual acuity in the information uptake process for the current simulations is an oversimplification of reality. Orthographic feature availability is an all-or-none phenomenon in the current version of the NDL model, with orthographic features being either present or absent in the input. Encoding feature availability over time as a gradient rather than an all-or-none phenomenon is likely to further improve the precision of the NDL model (see, e.g., Engbert et al., 2005). Nonetheless, the crude approximation of visual acuity in the information uptake process used here is an advancement over most existing models of compound processing that allowed for satisfactory performance of the NDL model when multiple fixations were required to process a compound.

At the second fixation, the eye moves further into the word, proportional to the compound's length. The primary predictor now, however, is the prior probability of the compound's lexeme. It seems likely that during the first fixation, a hypothesis space is set up that anticipates the compound lexemes with which the modifier lexeme co-occurs. At the second fixation, we observe that top-down priors are validated against the input. It is only when a compound is very long, and its lexemic prior extremely low, that a third fixation is required. The effect of the LSA similarity between the modifier and the compound on third fixation duration suggests that modifiers and compounds that have greater topical co-occurrence probabilities require less time to sort out how they contribute to the understanding of the discourse.

An important aspect of the discriminative approach to compound reading is that multiple lexemes can be co-activated. Baayen, Shaoul et al. (2015) work out this idea, taking as point of departure the point made by Ferdinand de Saussure that in language everything is interdependent.

De Saussure (1966) illustrated this point with an analogy to the game of chess. What a given piece contributes to the game is determined not only by what piece it is (a pawn or a rook), but also on where it is positioned on the board, and on the positions of the other pieces on the board. Thus, understanding a compound such as *wheelchair* requires knowledge of chairs, wheelchairs, and the kind of wheels one finds on wheelchairs (typically with a circular handgrip). From this perspective, the compound activation measure that predicts first-and-only fixation durations is of interest, as this measure is a weighted sum of the activations of modifier, head, and compound, suggesting that all three lexemes become available at the same time, and jointly drive the interpretation of the compound.[8]

When these three lexemes are considered as pieces on an interpretational chessboard, their mutual positions and strategic values are critical. For the head and compound lexemes at issue in our data, the compound lexeme typically discriminates a subclass of the experiences discriminated by the head. The experiences the head discriminates are to a considerable extent aligned. However, the modifier by itself discriminates very different experiences: the wheel in wheelchair is not a chair. Even though there is some systematicity in how specific modifiers relate to their heads, as laid out by the conceptual relations of Competition Among Relations In Nominals (CARIN) theory (Gagné & Shoben, 1997; Gagné & Spalding, 2014), this systematicity provides only the gist of an interpretation (e.g., not every chair that has wheels is a wheelchair). From this perspective, it makes sense that, to the extent that there is a signal in the noise, the LSA measure that appears to be predictive is the measure assessing modifier and compound similarity, rather than modifier and head similarity.

A chess-board contains much more than three pieces. Similarly, the mental lexicon contains thousands of lexemes. The approach described here is a first attempt to demonstrate that an interpretation of maximization of opportunities along the lines of "using every tool to open the lock" (i.e., to access the compound lexeme) is still an oversimplification

[8] Note that a similar integrative measure of the modifier, head and compound frequency provided substantially less explanatory power than the integrative NDL measure (REML scores for first-and-only fixation duration GAMMs: 414.05 versus 409.02).

of the real task at hand: generating a state of the full lexico-semantic system that allows for an adequate understanding of a compound word in the current linguistic context.

An interesting aspect of the current findings that we have not highlighted thus far is that the NDL analysis offers a more differentiated and more precise perspective on frequency effects. The NDL framework distinguishes two types of frequency effects. The first type of frequency effect reflects the amount of bottom-up support associated with a compound and/or its constituents and is captured by NDL activation measures from the orthography-to-lexeme network. This measure influences early measures of compound processing. The second type of frequency effect in the NDL model taps into top-down knowledge about the network connectivity of a compound through the MAD measure. This top-down information enters the equation when the bottom-up support for a compound is insufficient and additional fixations are required. In this case, readers fall back on a guessing strategy based on a priori knowledge about the prior probability of a compound, rather than attempting to process the compound in a bottom-up fashion once more.

In the lexical predictor analysis we observed a late emergence of semantic effects regarding the similarity of the modifier with the compound and the head. In the corresponding NDL models, we found a significant effect of the activation of the modifier on next fixation duration for first-and-only fixations, as well as a marginally significant effect of the activation of the modifier on third fixation duration ($F = 3.385$, $p = 0.067$). These effects, however, did not render the effects of the LSA similarity of the modifier with the compound and the head obsolete. In contrast to the effect of the semantic similarity of a compound to its head on first fixation durations, therefore, these late semantic effects cannot be explained through the bottom-up discriminability of a compound or its constituents.

Furthermore, the similarity measures derived from the lexeme-to-lexeme model that were included as systemic alternatives for the LSA measures did not reach significance for late measures of compound processing. In its current form, therefore, the NDL model proposed here

does not accurately capture the late semantic effects observed in the data. One potential explanation for this shortcoming may come from the size of the training corpus used here, the BNC. The median frequency of all 364 compound types in the BNC was no more than 100. As a result, the semantic similarity measures derived from the lexeme-to-lexeme network are based on a rather small number of training instances per compound. Consequently, the model may have been unable to accurately learn associations between a compound lexeme and all other lexemes in the lexicon.

Throughout this chapter we indicated that the NDL approach used here does not come without limitations. The all-or-none nature of the availability of orthographic information is a simplification of what is actually available to the eye. In addition, the NDL measures used in the current simulations are unable to capture both n-gram frequency effects and semantic similarity effects due to the relatively infrequent occurrence of compounds in the 100 million word BNC corpus on which the models were trained.

Nonetheless, the analyses presented here demonstrate that a discrimination learning approach to compound processing, in which readers simultaneously use all information available to them at a given point in time to generate a state of the lexico-semantic system that allows for efficient processing and an adequate understanding of the compound provides a highly competitive account of eye fixation patterns for the compounds in the ET corpus as compared to an extensive set of lexical predictors that underlie the architectures of existing sub-lexical, supra-lexical and dual route models of compound processing. As such, the informativeness of the NDL framework for the current data provides further evidence for the explanatory power of discrimination learning models of language processing - not only for reaction times in lexical decision experiments (Baayen et al., 2011, 2013; Ramscar et al., 2014) or word naming latencies (see Chapter 2), but also for the eye movements during compound reading in natural discourse contexts.

4

Picture naming

4.1 Introduction

Few effects in the psycholinguistic literature are better documented than the word frequency effect: the more often a word occurs in the language, the faster and more accurate people respond to that word in a wide range of linguistic tasks, including lexical decision (see, e.g., Scarborough et al., 1977; Balota et al., 2004) and word naming (see, e.g., Forster & Chambers, 1973; Balota & Chumbley, 1985; Jared, 2002). Recently, a number of studies have shown that word frequency effects are also present in electroencephalograms (EEGs) following the onset of a (linguistic) stimulus, which are commonly referred to as event-related potentials (ERPs).

Typically, the effects of word frequency on ERPs arise rapidly after the onset of the stimulus. Hauk et al. (2006), for instance, found an effect of word frequency in a visual lexical decision task as early 110 ms after stimulus onset. This early effect of word frequency was most prominent in left-lateralized temporal and parietal areas. Similarly, Sereno et al. (1998) found a word frequency effect in a visual lexical decision task that first reached significance at 132 ms after stimulus onset, whereas Penolazzi et al. (2007) observed an effect of word frequency in a sentence-reading task

that started at 120 ms after written word onset. The topographically widespread effect of word frequency in the picture naming task used by Strijkers et al. (2010) arose somewhat later, with more positive voltages for high frequency words than for low frequency word from 150 ms until voice onset.

The effect of frequency, however, is not limited to the word level. Arnon and Snider (2010) showed that phrasal decision latencies for high frequency phrases such as "all over the place" are shorter than those for low frequency phrases, such as "all over the city". This effect did not reduce to frequency effects of single words or smaller n-grams. The n-gram frequency effect has been replicated in a number of recent studies, showing n-gram frequency effects in sentence repetition (Bannard & Matthews, 2008), sentence reading (Siyanova-Chanturia et al., 2011), sentence recall (Tremblay et al., 2011) and frequency rating (Shaoul, Westbury & Baayen, 2013) tasks. Tremblay and Baayen (2010) added to these findings by observing an n-gram frequency effect in a free recall ERP study. The temporal onset of this effect was similar to that of the effects of word frequency described above, with n-gram probability first being significant around 110 ms after stimulus onset.

The n-gram frequency effect is theoretically interesting. At the very least, it "add[s] multi-word phrases to the units that influence processing in adults" (Arnon & Snider, 2010, p.76), which suggests that language users "seem to have [...] some experience-derived knowledge of specific four-word sequences" (Bannard & Matthews, 2008, p.246). Much, however, remains unclear about how this knowledge is implemented, and, therefore, about the implications of n-gram frequency effects for different models of language processing.

One interpretation of n-gram frequency effects is to consider these effects as evidence for whole-phrase representations. As noted by Baayen et al. (2013), such an interpretation fits well with theoretical approaches like data-oriented parsing (Bod, 2006) or memory-based learning (Daelemans & Bosch, 2005), in which large numbers of multiword sequences (or parse trees for these sequences) are stored in memory and optimal performance is ensured through on-line generalization over these stored sequences.

In exemplar-based approaches, therefore, *n*-gram frequency effects are directly related to the *n*-gram representations that are stored in memory.

Baayen et al. (2013), however, argued that storing each multiword sequence and its associated frequency in memory is associated with a number of problems. Given the Zipfian shape of frequency distributions, the number of unique *n*-grams is extremely large. The British National Corpus, for instance, contains 40 million unique word trigrams. Baayen et al. (2013) continue their argument by stating that even if the storage of gigantic numbers of word *n*-grams were neuro-biologically possible, on-line processing over an instance space of this size would be very time-consuming. To side-step this problem, the memory-based learning system implemented in TiMBL (Daelemans et al., 2007) uses information gain trees (Daelemans et al., 1997) as a compression algorithm to reduce the computational demands of on-line searches.

An additional problem with *n*-gram representations described by Baayen et al. (2013) is that it is not immediately clear what the function of such representations would be. Positing representations as a locus for a frequency "counter in the head" seems unconvincing (see, e.g., McClelland & Rumelhart, 1981 and Norris & McQueen, 2008 for models that integrate word unigram frequencies as a priori-probabilities). The application of shortlists in interactive activation models (Norris, 1994) raises further questions about the necessity of *n*-gram representations. These models use shortlists of stored candidates as a computational shortcut that allows for simulations with realistic input sizes. The success of shortlists in these types of models indicates that at least some stored multiword sequences are not relevant for on-line processing.

These concerns have led researchers to propose alternative explanations for the effect of *n*-gram frequency. Tremblay et al. (2011) suggest that *n*-gram frequency effects may reflect past experience with (de)compositional processing. Such an interpretation fits well with evidence from the learning literature demonstrating that "learning is a dynamic discriminative process" that is associative in nature (Ramscar et al., 2010). Baayen et al. (2013) argued that holistic representations may be beneficial at the earliest stages of learning (see, e.g., Dabrowska,

2000; Tomasello, 2003), but that additional experience will reduce the association strength between the linguistic components of these holistic initial representations and lead to an increased importance of decomposed, lower-level representations. Learning theory therefore predicts that the adult language processing system is less likely to have separate representations for multiword units (see Dabrowska, 2000; Arnon & Ramscar, 2012).

Baayen et al. (2013) provided computational support for such an interpretation of the n-gram frequency effect by successfully simulating the findings of Arnon and Snider (2010) in a full-decomposition model based on discrimination learning. The Naive Discriminative Reader (NDR) model used in their simulations has no representations beyond the simple word level. In the NDR model the n-gram frequency effect arises as a result of the associative learning process that maps orthographic input units (letters and letter combinations) to lexico-semantic units (word meanings). A high frequency phrase such as "all over the place" is read faster than a low frequency phrase such as "all over the city", because the letters and letter combinations in "all over the place" are more associated with the lexico-semantic representations *ALL*, *OVER*, *THE* and *PLACE* than the letters and letter combinations in "all over the city" are associated with the lexico-semantic representations *ALL*, *OVER*, *THE* and *CITY*.

Thus far we discussed effects of the frequency of multi-word sequences. The prototypicality of phrases is likewise reflected in behavioral measures of language processing. Several studies have documented prototypicality effects at the word level, using relative entropy to gauge the similarity of an exemplar to its constructional prototype (Milin, Filipović Durdević & Moscoso del Prado Martín, 2009; Milin, Kuperman et al., 2009; Kuperman et al., 2010). Above the word level, relative entropy effects have been observed for English prepositional phrases (Baayen et al., 2011). Given estimated probabilities p (relative frequencies) of prepositional phrases for a given noun and estimated probabilities q (relative frequencies) of prepositions across all nouns, prepositional relative entropy is defined as:

$$\text{Relative Entropy} = \sum_{i=1}^{n}(p_i * \log_2 (p_i/q_i)) \tag{4.1}$$

where n is the number of prepositions taken into account.

The relative entropy measure compares how similar the distribution of prepositional phrase frequencies for a given noun is to the distribution of preposition frequencies in the language as a whole. Values for relative entropy are low when the prepositional phrase frequency distribution for a given noun (exemplar) is similar to the overall prepositional phrase frequency distribution (prototype) and high when the prepositional phrase frequency distribution for a given noun differs substantially from the overall prepositional phrase frequency distribution. Higher relative entropies are typically associated with greater processing costs. Nouns that use prepositions in an atypical way, for instance, take longer to process than nouns that use prepositions in a typical way (Baayen et al., 2011). The effect of prepositional relative entropy implies that the language processing system is sensitive to the distributional properties of a noun's prepositional paradigm vis-a-vis the distribution of prepositional frequencies in the language as a whole.

From an exemplar-based point of view, the relative entropy measure may characterize part of the complexity of the exemplar space. How exactly this knowledge would be implemented in an exemplar-based model is unclear. One option would be on-line computation of the distance between the prepositional phrase frequency distribution for a given noun and the prepositional frequency distribution in the language as a whole. This, however, would involve tremendous amounts of online computation. Alternatively, the frequency distribution of the prototype (i.e., the frequency distribution of prepositions across all nouns) – or the distance between the frequency distributions for a given noun and the prototype – could be stored. This, however, would further increase the memory demands on the language processing system. In addition, it is unclear what function prototype distribution representations would have beyond accounting for the effect of relative entropy.

Discrimination learning offers an alternative explanation for relative entropy effects. Using a discrimination learning model without any representations beyond the basic word level, Baayen et al. (2011) successfully captured the fact that nouns with high prepositional relative entropies (i.e., nouns that use prepositions in an atypical way) take longer to process than nouns with low relative entropy. In the naive discrimination learning framework, the effect of relative entropy arises as a straightforward consequence of the way the distributional properties of English shape the associations between orthographic input cues and semantic outcomes across sequences of words.

4.2 Experiment

In what follows we present the results of a primed picture naming experiment that gauges the effects of word frequency, phrase frequency and phrase prototypicality using event-related potentials (ERPs). The current work seeks to extend previous findings in two ways. First, while previous studies have investigated the effects of word frequency on ERPs in a variety of tasks, the experimental results for phrase frequency and relative entropy discussed thus far were mostly obtained in chronometric studies. While these studies demonstrated that both frequency and relative entropy influence how (prepositional) phrases are processed, they offer little information on the temporal details of these effects. The temporal resolution of ERPs will allow us to gauge the millisecond-by-millisecond temporal development of the phrase frequency and relative entropy effects in a picture naming task. In addition, while the spatial resolution of ERPs is limited, the current work may provide us with a general idea about the topographical dynamics of these effects. The first goal of the current study, therefore, is to obtain a more detailed picture of the effects of word frequency, phrase frequency and relative entropy that arise during prepositional phrase processing.

The second goal of the current work is to find out to what extent measures derived from a naive discrimination learning model provide further insight into the temporal and spatial dynamics of the ERP signal

in a primed picture naming task. The discriminative learning approach has been shown to successfully simulate a variety of behavioral measures, including lexical decision latencies Baayen et al. (2011), word naming latencies (see Chapter 2) and eye movements during natural discourse reading (see Chapter 3). Predicting the ERP signal following the presentation of a prepositional phrase, however, involves predicting a signal as it evolves over both time and space. This stringent test of the discrimination learning approach will help gain more insight into the strengths and shortcomings of the discriminative learning approach to language processing.

In the present experiment, participants are presented with a preposition plus definite article prime, followed by a picture of a concrete noun that they have to name as fast and accurately as possible. The use of a primed picture naming paradigm offers a number of opportunities. First, prepositional relative entropy is a measure of constructional prototypicality: it describes how prototypical a given noun's use of prepositions is. The effect of relative entropy is best measured at the noun. In other tasks, such as sentence reading, the temporal onset of noun processing is hard to determine. The current primed picture naming task avoids this problem and precisely defines the earliest possible point in time where noun processing can take place as the moment the target noun picture appears on the screen.

A related benefit using a primed picture naming paradigm is that it reduces the temporal overlap between processes related to the preposition and definite article and processes related to the noun. Experienced readers are able to read prepositional phrases in a few hundred milliseconds. Nonetheless, as will become apparent soon, ERP effects related to the lexical properties of a given word can last many hundreds of milliseconds (see, e.g., Kryuchkova et al., 2012). This implies that there is a temporal overlap between processes related to the different words in the prepositional phrase. In the current setup, the temporal distance between the onset of the prime and the onset of the target is 2000 ms. This allows a substantial part of the initial processing of the preposition and definite article to complete prior to the presentation of the target noun.

A third reason for using the current experimental setup is that the proof of the pudding is in the eating as far as phrase frequency effects are concerned. As noted above, the current paradigm does not guarantee that the information in the preposition plus definite article primes is integrated with the information in the target noun picture to obtain a phrase-level understanding of the stimulus. Whether or not a phrase frequency effect can be observed in the primed picture naming task used here is an empirical issue. If we do observe an effect of phrase frequency, however, this unequivocally entails that the stimuli were processed at the phrase level.

The first part of what follows describes in more detail the experiment outlined above, the statistical methods used to analyze the data and the results of the experiment. In the second part, we present a simulation study in which we explore to what extent the discriminative learning framework can provide further insight into the temporal and spatial dynamics of the ERP signal following picture onset.

4.3 Methods

4.3.1 Participants

Thirty participants took part in the experiment. All participants were students of the University of Alberta in Edmonton and native speakers of English. Their mean age was 20.43 (sd: 4.67). Nineteen participants were female, eleven were male. All participants were right-handed, had normal or corrected to normal vision and did not have a history of neurological illness. Participants received partial course credit for their participation.

4.3.2 Materials

Sixty-eight concrete nouns were paired with photographs, depicting the referent of these nouns on a beige background. For each of the nouns, four three-word prepositional phrases were constructed, consisting of a preposition, the definite article "the" and the noun itself (e.g., "with the saw", "against the strawberry").

Phrases were selected on the basis of trigram frequencies as available in the Google 1*T* *n*-gram data (Brants & Franz, 2006). Trigram frequencies for all prepositional phrases consisting of a preposition, an article and one of the 68 concrete nouns were extracted. For a given noun, the phrases at 25%, 50%, 75% and 100% of the phrase frequency distributions were included as stimuli. For the noun "finger", for instance, this procedure generated the experimental items "over the finger" (25% of the phrase frequency distribution), "off the finger" (50%), "in the finger" (75%) and "with the finger" (100%). The total number of stimuli was 272.

Only prepositions from a pre-compiled list of 35 prepositions were included in the trigram frequency list. Selecting the phrases at the quantiles of the phrase frequency distribution led to 29 of these prepositions being used in the experiment. As a result of this selection procedure, there was a significant correlation between (log) preposition frequency and number of times a preposition was used in the experiment ($r = 0.85$, $p < 0.001$), with frequent prepositions such as "in" (44 times) or "on" (23 times) being included more often than infrequent prepositions such as "under" (6 times) or "against" (5 times). The experience with prepositions in the context of the current experiment therefore reflects the experience with prepositions in the language as a whole.

4.3.3 Design

The experiment consisted of 272 picture naming trials. Prior to the experiment, a practice phase was included, consisting of 10 items. The order in which the stimuli were presented was randomized between participants. The dependent variable was the ERP signal measured at 32 locations on the scalp. The independent variables were *Picture Complexity*, *Preposition Length*, *Word Length*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy*.

Picture Complexity is the size of the picture file in bytes. *Preposition Length* and *Word Length* are the length of the preposition and the target noun in letters. *Preposition Frequency*, *Word Frequency* and *Phrase Frequency* are the frequency of the preposition (e.g., "with"), target noun (e.g., "finger") and phrase (e.g., "with the finger") in the Google *n*-gram

data. *Picture Complexity*, *Preposition Length*, *Word Length*, *Preposition Frequency*, *Word Frequency* and *Phrase Frequency* were log-transformed prior to analysis to remove a rightward skew from the predictor value distribution. *Relative Entropy* was calculated on the basis of the Google n-gram phrase frequencies for prepositional phrases with definite article for all 68 nouns used in the experiment and all 35 prepositions in the precompiled list of prepositions. Prepositional phrase frequencies were converted to relative frequencies (i.e., estimated probabilities) for each noun and across all nouns to obtain estimated probability distributions p (for a given noun) and q (across all nouns). *Relative Entropy* was then calculated as the Kullback-Leibler divergence between p and q (see Equation 4.1).

Prior to analysis, we removed predictor outliers (i.e., predictor values further than two standard deviations from the mean) from the data. This resulted in the exclusion of 0.00% of predictor values for *Preposition Length*, 1.54% of predictor values for *Word Length*, 1.92% of all predictor values for *Preposition Frequency*, 4.62% of all predictor values for *Word Frequency*, 5.77% of all predictor values for *Phrase Frequency* and 4.62% of all predictor values for *Relative Entropy*. Table 4.1 shows the range and adjusted range for all independent variables. In addition, it presents the mean, median and standard deviation of the predictor distributions after outlier removal.

The resulting data set is characterized by a considerable amount of collinearity ($\kappa = 123.16$). *Word Frequency*, for instance, correlates positively with *Phrase Frequency* ($r = 0.42$) and negatively with *Preposition Frequency* ($r = -0.40$), *Relative Entropy* ($r = -0.40$) and *Word Length* ($r = -0.51$). Similarly, *Preposition Frequency* correlates not only with *Word Frequency*, but also shows a strong negative correlation with *Preposition Length* ($r = -0.76$).

One approach for dealing with collinearity is predictor residualization. In this approach, rather than entering the raw predictors into a regression model, one or more of the predictors are residualized prior to analysis by running a preliminary regression analysis with the predictor that is to be residualized as the dependent variable and one or more other predictors as

Table 4.1. Summary of the independent variables (*log*) *Picture Complexity*, (*log*) *Preposition Length*, (*log*) *Word Length*, (*log*) *Preposition Frequency*, (*log*) *Word Frequency*, (*log*) *Phrase Frequency* and *Relative Entropy*. Range is the original range of the predictor. Adjusted range is the range after removing predictor outliers. Mean, median and sd are the means, medians and standard deviations after outlier removal.

predictor	range	adj. range	mean	median	sd
Picture Complexity	8.53 - 11.13	8.69 - 10.83	9.88	9.91	0.50
Preposition Length	0.69 - 1.95	0.69 - 1.95	1.15	1.38	0.45
Word Length	1.10 - 2.30	1.10 - 2.08	1.58	1.61	0.26
Preposition Freq.	15.65 - 23.17	17.63 - 23.17	21.09	21.81	1.61
Word Frequency	12.90 - 18.96	13.60 - 18.37	15.74	15.50	1.25
Phrase Frequency	0.00 - 14.69	6.77 - 12.65	8.73	8.57	1.23
Relative Entropy	0.10 - 2.34	0.10 - 1.39	0.54	0.55	0.28

the independent variable(s). For the current data, for instance, it would be an option to residualize *Phrase Frequency* from *Word Length*, *Word Frequency*, *Preposition Frequency* and *Relative Entropy*. The resulting *Phrase Frequency* measure would then no longer correlate with these predictors.

Recently, however, Wurm and Fisicaro (2014) argued that residualization is not a useful remedy for collinearity. Contrary to popular believe, they state, residualization "does not change the results for the predictor that was residualized [... and ...] does not create an improved, purified, or corrected version of the original predictor" (Wurm & Fisicaro, 2014, p.45). What residualization does do, the authors continue, is introduce an additional statistical problem: depending on the correlation between predictor X_1 and predictor X_2 and the correlations between the dependent variable Y and predictors X_1 and X_2, residualization of X_1 results in either underestimating or overestimating the statistical importance of the non-residualized predictor X_2. Given these considerations, they conclude that, in the context of collinearity issues, "residualization of predictor variables is not the hoped-for panacea" (Wurm & Fisicaro, 2014, p.47).

Not all is bad, however. While suppression is a serious problem when it occurs, it may not be as common as previously thought. As noted by Wurm and Fisicaro (2014), for instance, Darlington (1990, p.155) states that "suppression rarely occurs in real data", and J. Cohen et al. (2003) argues that "it is more likely to be seen in fields like economics, where variables or actions often have simultaneous equilibrium-promoting effects". Although the correlation threshold for potential suppression depends on the correlation of the involved predictor with the dependent variable, suppression artifacts are highly uncommon for weak or moderate correlations.

For the current data set, these statements suggests that while suppression is not outside the realm of possibilities for the effects of *Preposition Length* and *Preposition Frequency*, our analysis of the main predictors of interest (*Word Frequency*, *Phrase Frequency* and *Relative Entropy*) is unlikely to suffer from this problem. We therefore decided to use the original, non-residualized measures *Picture Complexity*, *Preposition*

Length, *Word Length*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* described above as predictors in our analysis.

To ensure that the results from this analysis were robust, we carried out a post-hoc analysis. For each of the predictor effects reported below, we fitted a new model with an identical model structure, but omitting the other lexical predictors. The results of this post-hoc analysis were qualitatively similar to the effects reported below. The problem of suppression therefore appears to be limited for the current data set.

4.3.4 Procedure

Data were recorded from 32 Ag/AgCl active electrodes (*Fp1*, *Fp2*, *AF3*, *AF4*, *F7*, *F3*, *Fz*, *F4*, *F8*, *FC5*, *FC1*, *FC2*, *FC6*, *T7*, *C3*, *Cz*, *C4*, *T8*, *CP5*, *CP1*, *CP2*, *CP6*, *P7*, *P3*, *Pz*, *P4*, *P8*, *PO3*, *PO4*, *O1*, *Oz*, *O2*), which were mounted on an electrode cap (BioSemi, international 10/20 system). Reference electrodes were placed at the left and right mastoids. The EOG was recorded using electrodes below and above the left eye and at the outer canthi of both eyes. Electrode cap sizes varied from 54 to 60 cm between participants to allow for an optimal fit.

Data were sampled at $8,102$ *Hz* using a BioSemi Active II amplification system. Prior to analysis, the signal was down-sampled to 256 *Hz*, band-pass filtered from 0.5 to 50 *Hz*, baseline corrected (-200 to 0 ms interval) and re-referenced to the average of the left and right mastoids using Brain Vision Analyzer (version 1.05). In addition, the signal was corrected for eye-movements and eye blinks using the *icaOcularCorrection* package for R (Tremblay, 2010).

Verbal responses were recorded using a microphone (Sennheiser) and response box including a voice key (Serial Response Box) for the E-Prime experimental software package (version 2.0.1). The same package was used to present the stimuli on a 17 inch CRT monitor using a 1024 by 768 resolution.

A fixation mark was shown for 1000 ms prior to each trial. Next, participants were presented with a preposition plus definite article prime (e.g., "in the") for 1000 ms. This screen was followed by another 1000 ms

fixation mark screen. We then presented the photograph depicting the target noun (512 by 384 pixels) for 3000 ms. Participants were instructed to name the target noun, as depicted by the photograph. They were asked to respond as fast a possible, while retaining accuracy. In addition, participants were instructed to limit eye blinking and body movements to a minimum.

All fixation marks and texts were presented in white Courier New 24 point font. All fixation marks, texts and photographs were presented in the center of the screen against a black background. Each photograph was followed by a 2000 ms pause to allow the EEG signal to return to baseline prior to the next stimulus. The experiment had a duration of about 40 minutes, excluding a preparation phase of about 30 minutes. Halfway through the experiment, participants were given a break to prevent fatigue.

4.4 Analysis

Prior to analysis we removed 12 items (4.41%) corresponding to 3 problematic photographs from the data, as error rates were high for these photographs across participants. In addition, we removed incorrect naming responses from the data (2.79%). Trials for which the maximum absolute voltage after signal correction exceeded 100 μV at any channel were removed from the data for all channels (5.25%). Furthermore, 39 trials (0.48%) were removed due to technical failure. This resulted in a total data loss of 12.93%. No averaging over participants or items was done prior to analysis.

4.4.1 Generalized Additive Models (GAMs)

This experiment examines the effect of numerical predictors over time. These effects are potentially non-linear in both the predictor dimension (at a given point in time) and the time dimension (for a given predictor value). To allow for non-linearities in multiple dimensions, we used generalized additive mixed-effect models (GAMMs; Hastie & Tibshirani, 1986; Wood, 2006) as implemented in the R package *mgcv* (version 1.8.3)

to analyze our data. GAMMs have recently been used in a number of ERP studies on language processing (Kryuchkova et al., 2012; Baayen, Tremblay & Hendrix, 2015; Hendrix, 2008).

The use of regression models has become commonplace in experimental studies investigating predictor effects on unidimensional dependent variables, such as reaction time studies. The application of regression type models in ERP studies, however, is much less widespread. To allow for a better understanding of the analysis technique used here and the advantages GAMMs offer in comparison to a traditional ERP analysis we compare the current ERP analysis to a traditional ERP analysis for simulated data, as well as for some of the key predictor effects described below in Appendix A.

4.4.2 Reaction time analysis

We fitted a GAMM with by-participant factor smooths for trial and a random intercept for noun (e.g., "finger") to the naming latency data. Random intercepts for preposition (e.g., "with") and phrase (e.g., "with the finger") were not significant and therefore omitted from the reported model. Naming latencies further than 2 standard deviations from the mean were removed from the data prior to analysis. A log transformation was applied to the naming latencies to remove a rightward skew from the naming latency distribution. We modeled the predictor effects of *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* using smooth functions ($k = 5$). The effects of *Preposition Length* and *Word Length* were modeled with a parametric term, because of the limited number of unique values for these predictors. As such, linearity was imposed for the effects of *Preposition Length* and *Word Length*.

4.4.3 ERP analysis

For each electrode, we fitted a GAMM with a main effect smooth for time, by-participant factor smooths for time and trial, as well as random intercepts for preposition, noun and prepositional phrase to the ERP signal

from 0 to 600 ms after picture onset. For each of the predictors *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* we furthermore included a main effect smooth, as well as a tensor product interaction with time (modeled through `ti()` terms). We also included main effect smooths for *Preposition Length* and *Word Length*. These main effect smooths for *Word Length* and *Preposition Length*, however, reached significance at 1 electrode only (*Word Length*: electrode *C4*, $p = 0.023$; *Preposition Length*: electrode *AF4*, $p = 0.020$). Given the issue of multiple comparisons, these results provide little evidence for a statistically robust effect of *Word Length* or *Preposition Length*. We therefore decided not to include the main effect smooths for *Preposition Length* and *Word Length* in the GAMMs reported in this chapter. Effects in the predictor dimension were limited to 5th order non-linearities ($k = 5$), whereas effects in the time dimension were to 20th order non-linearities ($k = 20$). To control for AR1 autocorrelation processes, we included an autocorrelation parameter ρ in the GAMMs, which was set to 0.75.

Figure 4.1 shows the predicted values of our GAMM at electrode *C3* (black line). Predicted main trend values correlate highly with average observed voltages (red dots): $r = 0.999$. This indicates that the GAMM successfully captures the general trend of the ERPs over time. GAMM fits correlated highly with averaged observed voltages across all electrodes, with an average correlation of $r = 0.997$ between predicted values and average observed values.

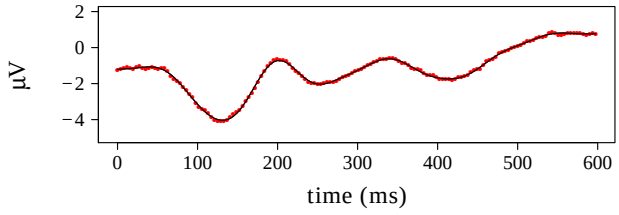

Figure 4.1. Main trend in the ERP signal at electrode *C3* as predicted by the main trend GAMM (black line) and as observed (red dots).

The average reaction time in the experiment was 854 ms (median: 800 ms). The earliest responses started coming in much earlier than that. As can been seen in the left panel of Figure 4.2, articulation has begun for a significant proportion of trials by the end of our 600 ms analysis window (13.6%). As a consequence, electromyographic (EMG) potentials arising from the facial, jaw and tongue muscles are present in a substantial subset of our data. These EMG potentials could therefore impoverish the signal-to-noise ratio (SNR) for this subset of the data.

There are two options for dealing with EMG activity in our data. First, we could remove all data points after the onset of articulation. As noted by Hillyard and Picton (1987), however, muscle artifacts may well be present long before speech onset. Even if we were to remove all data points following the onset of articulation, EMG artifacts would therefore remain in the data. Second, as noted above, articulation has started for 13.6% of all trials before the end of the 600 ms analysis window. Furthermore, the voice key did not register naming latencies for a non-trivial number of trials (for details, see the reaction time results section). Given that we are unsure about whether or not articulation started before the end of our analysis window, we would have to exclude these trials entirely avoid articulation artifacts altogether. Removing these data points and trials from the analysis would result in a substantial loss of statistical power.

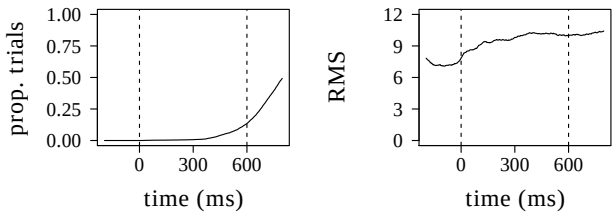

Figure 4.2. Left panel: proportion of data points after the onset of articulation as a function of time. Right panel: average root mean square (RMS) of μV across all electrodes from -200 to 800 ms after picture onset (0 ms).

The second option for dealing with EMG activity is to include all data points, even those for which articulation artifacts might be present. While this approach ensures an equal amount of data for each point in time, it does not necessarily solve the problem of reduced statistical power after the onset of pronunciations. If EMG artifacts have a negative effect on the SNR after pronunciation onset it becomes harder for statistical models to identify predictor effects. To gauge the severity of this problem, we calculated the root mean square (RMS) for all electrodes. The right panel of Figure 4.2 shows the average RMS across all electrodes as a function of time. In the pre-stimulus interval (−200 to 0 ms), the average RMS across all electrodes and time points is 7.31, whereas in the post-stimulus interval (0 to 600 ms) it is 9.96. As predicted, the RMS does increase as a function of time. The increase, however, is fairly limited: the average RMS is 8.98 in the 0-200 ms interval, 9.83 in the 200-400 ms interval and 10.13 in the 400-600 ms interval. Furthermore, the increase in RMS primarily occurs in the first 400 ms after picture onset, but stabilizes in the 400-600 ms time window. Given that only 2.11% of the articulations began prior to the 400 ms mark, the early increase in RMS values is unlikely to be due to muscle artifacts following the onset of articulation.

To further inspect the potential problem of a decreased SNR due to articulation artifacts we looked at the SNR across electrodes in the last 200 ms of our analysis window (i.e., 400-600 ms after picture onset). If articulation introduces noise in the signal, we would expect this noise to be most prominent at frontal electrodes, which are closest to the facial, jaw and tongue muscles. RMS averages in the last 200 ms were indeed elevated at frontal locations. While the average RMS across all electrodes in the 400-600 ms time window was 10.13, the average RMS values in this epoch at frontal electrodes were 15.02 (*Fp1*), 14.01 (*Fp2*), 13.13 (*AF3*), 11.67 (*AF4*), 12.51 (*F7*), 11.66 (*F3*), 8.62 (*Fz*), 9.72 (*F4*), 12.10 (*F8*), 10.32 (*FC5*), 10.34 (*FC1*), 6.51 (*FC2*) and 9.50 (*FC6*). As such, the average RMS values at frontal electrodes show an increase in the last 200 ms. This increase, however, is limited to the most frontal electrodes only.

Despite the topographically limited and quantitatively moderate increase in RMS values over time, articulation artifacts could nonetheless be problematic if they vary systematically with our predictors of interest. To rule out this possibility, we compared the results of an analysis on the full data set to the results of an analysis on a subset of the data that excluded all trials with naming latencies shorter than 600 ms, as well as trials for which no naming latencies were available. As such, this analysis excluded all potential muscle artifacts following articulation onset. The results of this analysis were highly similar to the results of the analysis on the full data set. We therefore decided to carry out our analysis on the full data set, including data points after articulation onset and trials for which no naming latencies were available.

4.5 Reaction time results

During the experiment there were some technical difficulties regarding the sensitivity of the voice key. This resulted in response times not being registered for 2 participants. These participants therefore could not be included in the reaction time analysis. In addition, we removed all further trials for which the voice key did not register a response (7.82%) from the data prior to the reaction time analysis

The naming latencies showed a significant random intercept for the target noun ($F = 11.614$, $p < 0.001$), and significant by-participant factor smooths for trial ($F = 12.831$, $p < 0.001$). Furthermore, we observed a significant effect of *Picture Complexity* ($F = 3.807$, $p = 0.017$). The effect of *Picture Complexity* is depicted in Figure 4.3. For ease of interpretation, predicted linear naming latencies are plotted rather than the log transformed latencies used for modeling (through adding the model intercept to the partial effect of *Picture Complexity* and applying an exponential transformation before plotting).

As can be seen in Figure 4.3, the effect of *Picture Complexity* is quadratic in nature, with low *Picture Complexity* leading to longer naming latencies and the effect leveling off for high predictor values. This effect of *Picture Complexity* is perhaps most easily interpreted by taking into

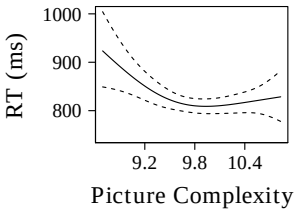

Figure 4.3. Effect for (log) Picture Complexity on the naming latencies.

consideration that *Picture Complexity* is proportional to information: the more complex a picture, the more information it contains. The longer naming latencies for pictures with limited complexity, therefore, may be a result of the fact that less complicated pictures do not contain enough information for a rapid identification of the depicted object. No other lexical predictors had a significant effect on the naming latencies.

4.6 ERP results

In this section, we discuss the results for the predictors *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy*. For each predictor, we visualize the partial effect of the time by predictor tensor product, as well as the main effect over time at a representative example electrode. Given the fact that GAMMs tend to be somewhat unreliable near the edges, we selected representative example electrodes that did not display potentially unreliable behavior near the edges of the analysis window whenever possible.

4.6.1 Picture Complexity

Figure 4.4 shows the contour plot of the partial effect of the tensor product interaction between time and *Picture Complexity*. The x-axis represents time (in ms) at a representative example electrode. *Picture Complexity* is on the y-axis. The contour plot represents voltages at the depicted electrode, with warmer colors representing higher voltages.

Contour lines are shown at intervals of 0.2 μV. The p-value for the effect at the depicted electrode is presented in brackets in the figure title.

Figure 4.4 furthermore contains a picture inset. This picture inset shows the topography of the effect, with dark red indicating significance at an alpha level of 0.05 and bright red indicating significance at a Bonferroni-corrected alpha level of (0.05/32 =) 0.0016. As can be seen in the inset in Figure 4.4, the tensor product between time and *Picture Complexity* is highly significant for a large number of electrodes across the scalp. A visual inspection of the results, however, revealed that the effect is most prominent in left and central parietal and occipital regions.

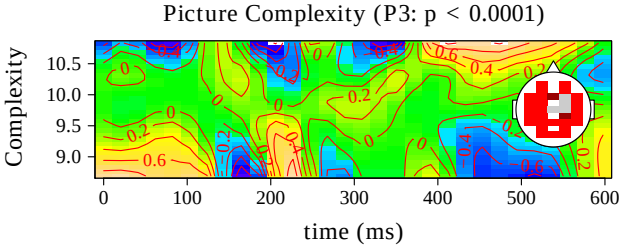

Figure 4.4. Effect for the tensor product interaction between time and (*log*) *Picture Complexity* at electrode *P3*. Color coding indicates voltages (in μV), with warmer colors representing higher voltages. Picture insets show the topography of the effect, with bright red indicating significance at the Bonferroni-corrected alpha level ($p < 0.0016$) and dark red indicating significance at the non-corrected alpha level ($p < 0.05$).

For both high and low values of *Picture Complexity*, Figure 4.4 shows that voltages are negative, then positive, then negative, then positive, et cetera. In other words, oscillations tied to the complexity of the presented picture are present in the ERP following picture onset. These oscillations have an opposite phase for low and high values of *Picture Complexity*: when complex pictures show high voltages, less complex pictures show low voltages and vice versa. To determine the frequency of the oscillations, we converted the time domain representation of the ERP signal seen in Figure 4.4 to the frequency domain. Although the frequency of the oscillations varies with time and predictor values, a peak

in spectral intensity that corresponds to the early oscillations for highly complex pictures and the oscillations for pictures with low complexity in the middle of the analysis widow is reached at 7 *Hz*. As such, these oscillations tied to *Picture Complexity* are in the upper part of the theta range (3 to 7.5 *Hz*).

To gauge the temporal onset of time by predictor tensor products, we calculated three sigma (99.7%) confidence intervals around the contour surfaces. The first point in time at which 0 is not within this three sigma confidence interval for high values of *Picture Complexity* is 46 ms after picture onset. The early positive voltages for low values of Picture Complexity, however, are already significant right after picture onset.[1] One potential explanation for the extremely early effect of *Picture Complexity* is that GAMM estimates can be somewhat unreliable near the edges of the analysis window. It could be the case that uncertainty about the effect for low complexity pictures in the first 50 ms led to a temporal overestimation of a positivity that started somewhat later in time. In the context of the expectations set up by the preposition plus definite article prime, however, the possibility of very early anticipatory responses to simple pictures can not be ruled out.

For each predictor we also fit a main effect smooth. The partial effect of this smooth term for *Picture Complexity* is presented in Figure 4.5, which shows that voltages seem to be somewhat increased for pictures with a higher visual complexity as compared to pictures with a lower visual complexity. As can be seen in the picture inset in Figure 4.5, however, the evidence for such an effect is anything but convincing: the main effect smooth of *Picture Complexity* reaches significance at a non-corrected alpha level at 2 electrodes only.

[1] Note that for oscillatory effects the phase of an oscillation co-determines the significance of an effect at a given point in time. Potential oscillations in the predictor dimension further complicate the process of determining the exact onset of an effect. As a result, the numbers reported for oscillatory effects here are conservative estimates for the temporal onset of these effects. In addition, as a result of phase shifts across the scalp these estimates are sensitive to the choice of the example electrode.

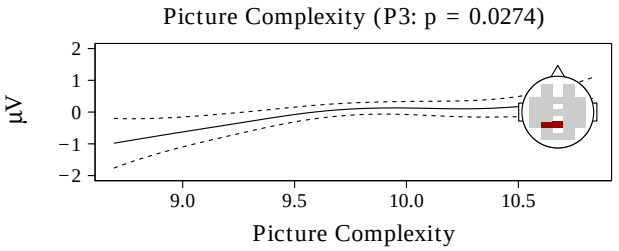

Figure 4.5. Effect for the main effect smooth of (*log*) *Picture Complexity* over time at electrode *P3*. Picture insets show the topography of the effect, with bright red indicating significance at the Bonferroni-corrected alpha level ($p < 0.0016$) and dark red indicating significance at the non-corrected alpha level ($p < 0.05$).

4.6.2 Preposition Frequency

Figure 4.6 presents the tensor product interaction of time by *Preposition Frequency*. The effect of *Preposition Frequency* is most prominent for low predictor values, with higher voltages for low frequency prepositions as compared to higher frequency preposition in the first 200 ms after picture onset. The fact that we see a significant effect of *Preposition Frequency* right after picture onset is unsurprising, given the fact that prepositions temporally preceded pictures in the experimental paradigm adopted here.

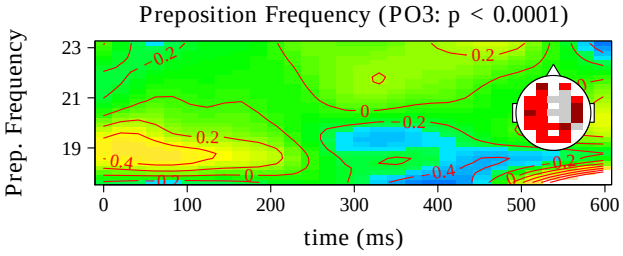

Figure 4.6. Effect for the tensor product interaction between time and (*log*) *Preposition Frequency* at electrode *PO3*.

After about 300 ms, the effect of *Preposition Frequency* reverses, with lower voltages for low frequency prepositions as compared to high frequency prepositions starting from 300 ms after picture onset. The effect of *Preposition Frequency* is topographically widespread, but more prominent in the left hemisphere than in the right hemisphere. The greatest effect sizes, however, were observed at left-lateralized parietal electrodes and bilateral occipital electrodes.

As for *Picture Complexity*, the results for the main effect smooth of *Preposition Frequency* showed little evidence for a *Preposition Frequency* effect over time. As can be seen in Figure 4.7, we found an effect at 2 electrodes at a non-corrected alpha level only, with slightly higher voltages for high frequency prepositions than for low frequency prepositions. As such, the effect of *Preposition Frequency* is much better described by a time by predictor interaction than by a main effect smooth.

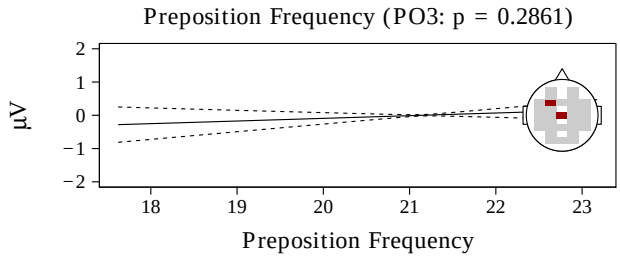

Figure 4.7. Effect for the main effect smooth of (*log*) *Preposition Frequency* over time at electrode *PO3*.

4.6.3 Word Frequency

Figure 4.8 shows the results for the time by *Word Frequency* tensor product interaction. The effect is characterized by oscillations for both high and low frequency words that are in opposite phrase and that reach maximum spectral intensity at 3 *Hz*. As such, these oscillations can be characterized as oscillations at the lower end of the theta range. Previously, theta range activity has been observed in a number of language processing studies and has been demonstrated to be related to,

for instance, lexical-semantic retrieval (Bastiaansen et al., 2005, 2008), syntactic processing (Bastiaansen et al., 2002) and translation (Grabner et al., 2007). In a regression study using GAMMs, Kryuchkova et al. (2012) recently reported theta range oscillations in auditory comprehension tied to word frequency, phonological neighborhood density and morphological family size. Theta range oscillations are thought to reflect (working) memory demands in language processing that arise from the synchronous firing of neurons in hippocampal areas (see Bastiaansen & Hagoort, 2003 for a comprehensive discussion of theta range oscillations).

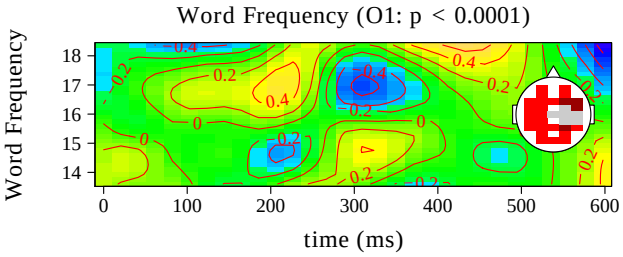

Figure 4.8. Effect for the tensor product interaction between time and (*log*) *Word Frequency* at electrode *O1*.

The effect of *Word Frequency* arises early. It is first significant at 95 ms after picture onset for medium to high predictor values. The early onset of the frequency effect for high frequency words is in line with previous findings (Hauk et al., 2006; Penolazzi et al., 2007; Sereno et al., 1998), reporting frequency effects in visual word recognition that arise between 110 and 132 ms after word onset. The oscillations for low frequency words are somewhat more subtle in nature than those for high frequency words, with smaller amplitudes and a later onset (for low frequency words, the effect of word frequency first reaches significance at 183 ms after picture onset).

The time by *Word Frequency* tensor product is significant at a large number of electrodes, with robust effects across frontal-to-occipital electrodes in the left hemisphere. By contrast, we found little to no evidence for a main effect of *Word Frequency* over time. Figure 4.9 shows that the

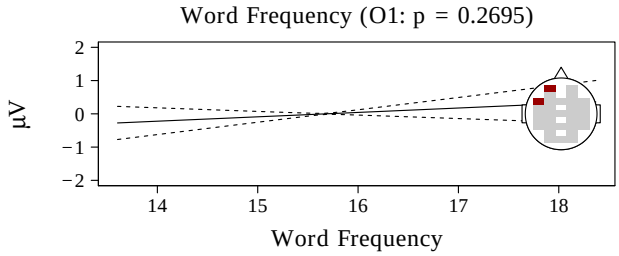

Figure 4.9. Effect for the main effect smooth of (*log*) *Word Frequency* over time at electrode *O1*.

main effect smooth for *Word Frequency* was significant at a non-corrected alpha level at 2 of the most frontal electrodes only. At these electrodes, we observed a small increase in voltages for higher values of *Word Frequency*, similar to the non-significant effect depicted in Figure 4.9 for electrode *O1*. As for the effect of *Preposition Frequency*, therefore, the effect of *Word Frequency* is much better described by a time by predictor interaction than by a main effect smooth.

4.6.4 Phrase Frequency

Figure 4.10 shows the tensor product interaction of time by *Phrase Frequency*. At first glance, it seems like there is a strong early positivity for high frequency phrases and a less pronounced early negativity for

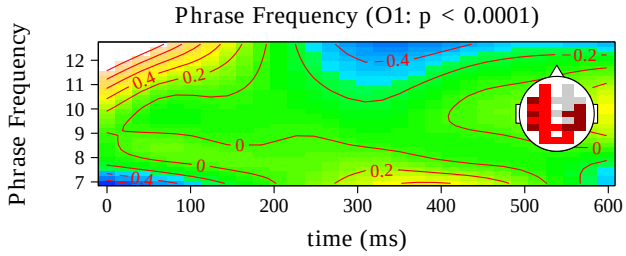

Figure 4.10. Effect for the tensor product interaction between time and (*log*) *Phrase Frequency* at electrode *O1*.

low frequency phrases, followed by a reversal of this pattern, with later negative voltages for high frequency phrases and positive voltages for low frequency phrases.

The main effect smooth of *Phrase Frequency*, however, reveals further insight into the tensor product interaction of time by *Phrase Frequency*. This main effect is presented in Figure 4.11. In comparison to *Preposition Frequency* and *Word Frequency*, *Phrase Frequency* shows more evidence for a main effect over time, with lower voltages for high frequency phrases as compared to low frequency phrases at a number of electrodes across the left hemisphere. The effect, however, reaches significance at a Bonferroni-corrected alpha level at 2 electrodes only. The evidence for a main effect over time for *Phrase Frequency*, therefore, is less than overwhelming.

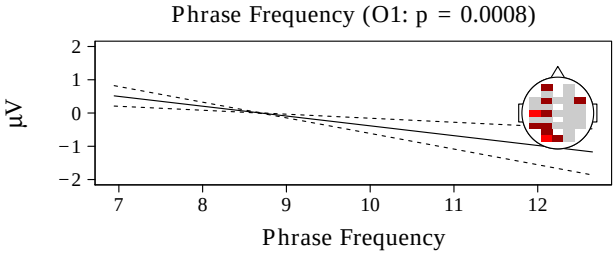

Figure 4.11. Effect for the main effect smooth of (*log*) *Phrase Frequency* over time at electrode *O1*.

As can be seen in Figures 4.10 and 4.11, the pattern of results for the time by *Phrase Frequency* interaction at the start of the analysis window is opposite to the main effect of *Phrase Frequency* over time, such that the main effect of *Phrase Frequency* is initially cancelled out by the time by *Phrase Frequency* interaction. To illustrate this point, Figure 4.12 presents the additive contour surface (i.e., the sum of the partial effect plots) for the main effect of *Phrase Frequency* (Figure 4.11) and the tensor product interaction between time and *Phrase Frequency* (Figure 4.10).

Figure 4.12 shows that the effect of *Phrase Frequency* is best characterized as a near-linear effect, with more positive voltages for low frequency phrases and more negative voltages for high frequency phrases. This effect arises somewhat earlier for low frequency phrases than for high frequency phrases and continues throughout the 600 ms analysis window. As such, the effect of *Phrase Frequency* is qualitatively different from the effect of *Word Frequency*, which was characterized by theta range oscillations, rather than prolonged effects over time.

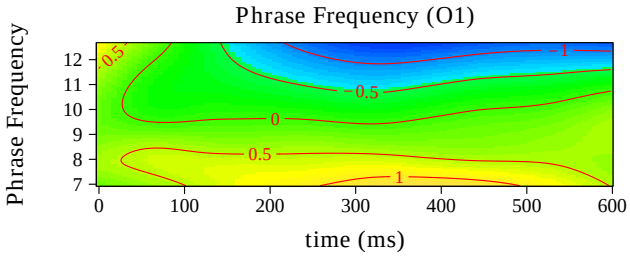

Figure 4.12. Additive contour surface for the tensor product interaction between time and (*log*) *Phrase Frequency* (Figure 4.10) and the main effect of (*log*) *Phrase Frequency* over time (Figure 4.11) at electrode *O1*.

4.6.5 Relative Entropy

Figure 4.13 presents the tensor product interaction between time and *Relative Entropy*. Similar to the effect of *Word Frequency*, the effect of *Relative Entropy* is characterized by theta range oscillations (4 *Hz*). These oscillations are most prominent for high values of *Relative Entropy*, although opposite-phase oscillations with a lower amplitude are present for medium-to-low values of *Relative Entropy* as well.

The effect of the tensor product interaction of time by *Relative Entropy* is topographically widespread, with significant effects across the left - and to a lesser extent - the right hemisphere. The effect is most prominent at parietal and occipital electrodes. For high values of *Relative Entropy*, the effect is first significant at 95 ms after picture onset, whereas for medium-to-low values of *Relative Entropy* the effect first reaches

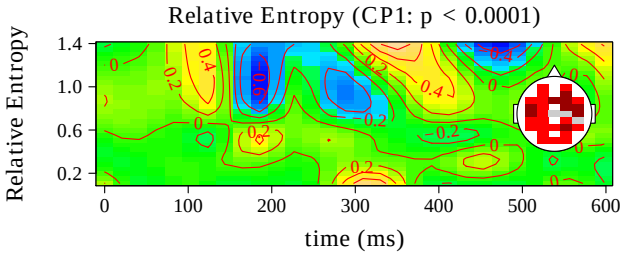

Figure 4.13. Effect for the tensor product interaction between time and *Relative Entropy* at electrode *CP1*.

significance at 104 ms after picture onset. As such, the temporal onset of the *Relative Entropy* effect is highly similar to that of the *Word Frequency* effect.

Reaction time studies reported increased response latencies for words with high relative entropies (Milin, Filipović Durdević & Moscoso del Prado Martín, 2009; Milin, Kuperman et al., 2009; Kuperman et al., 2010; Baayen et al., 2011). The current pattern of results fits well with these findings if we interpret the increased amplitude of the oscillations for high values of *Relative Entropy* as evidence for increased processing costs. The current results then indicate that additional processing is required for nouns with atypical prepositional phrase frequency distributions as compared to nouns that use prepositions in a more typical way.

For completeness, we conclude with the main effect smooth of *Relative Entropy*. As can be seen in Figure 4.14, we found little evidence for an effect of *Relative Entropy* over time. An effect at a non-corrected alpha level was found at 2 electrodes only, with somewhat decreased voltages for higher values of *Relative Entropy*. As for the effects of *Preposition Frequency* and *Word Frequency*, however, it is clear that the effect of *Relative Entropy* is best described by a tensor product interaction of time by *Relative Entropy*.

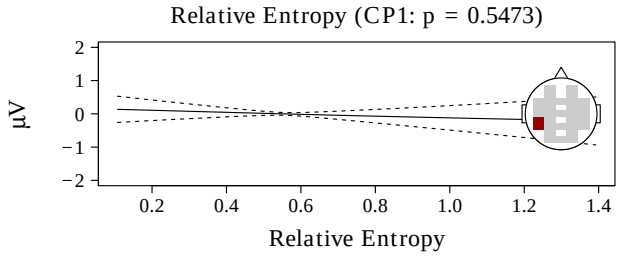

Figure 4.14. Effect for the main effect smooth of *Relative Entropy* over time at electrode *CP1*.

4.7 Discussion

In the current experiment, we observed effects of both word-level and phrase-level predictors in a primed picture naming paradigm. The effects of *Relative Entropy* and *Word Frequency* showed remarkable similarities. Both effects are characterized by oscillations at the lower end of the theta range. In addition, both effects showed similar topographical distributions and increased effect sizes in the left hemisphere as compared to the right hemisphere. Furthermore, the temporal onset of the effects was similar, with the onset of both effects being no more than 2 ms apart (*Word Frequency*: 97 ms after picture onset, *Relative Entropy*: 95 ms after picture onset). Neither *Word Frequency*, nor *Relative Entropy* showed a statistically robust main effect over time.

Similar to the effects of the word-level predictors *Word Frequency* and *Relative Entropy*, the effect for the phrase-level predictor *Phrase Frequency* was most prominent in the left hemisphere. In contrast to the effects of these word-level predictors, however, the effect for *Phrase Frequency* was not characterized by theta range oscillations. Instead, we observed a prolonged near-linear effect, with more negative voltages for high frequency phrases as compared to low frequency phrases. How should we interpret this pattern of results?

Exemplar-based models correctly predict the presence of both word frequency and phrase frequency effects in the current experiment. In exemplar-based approaches such as data-oriented parsing (Bod, 2006) or memory-based learning (Daelemans & Bosch, 2005), phrase frequency effects are explained through the existence of phrase representations (see Baayen et al., 2013). The frequency count associated with a phrase representation determines how quickly that phrase representation can be accessed, just like the frequency count associated with a word representation determines how quickly that word can be accessed.

Exemplar-based models are appealing because they account for phrase frequency effects in a very straightforward manner. Nonetheless, these types of models are associated with a few concerns in the context of phrase frequency effects. First, if similar dedicated representations underlie both word and phrase frequency effects, it is not immediately clear why the effects of word and phrase frequency observed here are qualitatively very different.

Second, as demonstrated by Baayen et al. (2013), exemplar sets very quickly become very large when dedicated n-gram representations are assumed. As a result, online processing over these exemplar sets would be very time-consuming. This concern is also relevant in the context of the relative entropy effect observed here. From an exemplar-based perspective, the relative entropy measure may describe how complex the exemplar space is. The fact that relative entropy shows robust effects on the ERP signal as it develops over time indicates that this knowledge is available to language users. How exactly this works, however, is unclear. One possibility is online-computation over the set of prepositional phrase exemplars. Given the size of this exemplar set, however, this would be a computationally expensive option. Another possibility is that (the distance between the prepositional phrase frequency distribution for a given noun and) the prototype distribution is stored in a set of "counters in the head". This, however, would put further demands on the storage capacity of the language processing system.

Although the current results are not incompatible with an exemplar-based approach, therefore, it is worth exploring alternative approaches that might offer more parsimonious accounts of the current data. Discrimination learning provides an alternative account for the effects of word frequency, phrase frequency and relative entropy that does not assume representations beyond the simple word level. Baayen et al. (2011) successfully replicated chronometric effects of prepositional relative entropy and phrase frequency in the Naive Discriminative Reader (NDR) model. In what follows, we explore to what extent naive discrimination learning (NDL) measures can provide further insight into the ERP signal in the current primed picture naming study. First, we describe the details of an NDL analysis on the basis of 4 predictors derived from two discrimination learning networks. Next, we present the results of this analysis for each of these discrimination learning measures.

4.8 NDL simulation

The NDL network in Baayen et al. (2011) maps orthographic units onto lexemes using a single discrimination learning network. Measures derived from this network afford good simulation results for silent reading. The task in the current experiment involves much more than silent reading. The orthographic presentation of the preposition and definite article is in line with the nature of the orthography-to-lexeme network in Baayen et al. (2011). By contrast, the target noun is depicted in a photograph. One option, therefore, would be to implement an additional discrimination learning network mapping visual features of the photograph onto the word meaning of the target noun. The implementation of such a network, however, is far from trivial. Furthermore, the focus of this simulation is on gauging the explanatory power of lexical network learning, rather than on describing the processes by which visual stimuli are recognized.

A second discrepancy between the current experimental setup and the orthography-to-lexeme network described in Baayen et al. (2011) concerns the nature of the task. While the orthography-to-lexeme network provides a silent reading model, the task in the current experiment

involves naming the target noun. In Chapter 2, we implemented the NDR$_a$ model, an extension of the original NDR model in Baayen et al. (2011) for reading aloud. The NDR$_a$ consists of two networks: a network mapping orthographic cues onto lexemes and a network mapping lexemes onto acoustic features (demi-syllables). The NDR$_a$ correctly predicts a number of findings that are specific to the reading aloud literature, such as effects of the consistency of the orthography to phonology mapping and a pseudo-homophone advantage for non-words.

Nonetheless, we decided to use a simple orthography-to-lexeme network in the current simulation for two reasons. First, the current task is somewhat of a hybrid between production and comprehension. At the word level, the task very much resembles a reading aloud task, albeit with visual rather than orthographic input. At the phrase level, however, no overt response is required. The effect of phrase frequency is an effect of implicit phrase-level comprehension, not of phrase-level production. While ideal for word-level simulations, therefore, the architecture of the NDR$_a$ is less than optimal for phrase-level simulations.

Second, despite the fact that the orthography to phonology mapping in English is inconsistent at times, there is considerable isomorphism between the orthographic and the phonological representations of words. As a result, there is a fair amount of overlap between the information learned by a discriminative learning network from orthography to semantics and the information learned by a discriminative learning network from phonology to semantics. For the set of $2,524$ monosyllabic words used in Chapter 2 for instance, the (log and inverse transformed) activation of the target word meaning from the orthography is highly correlated with the (log and inverse transformed) activation of the target word meaning from the phonology ($r = 0.48$, $p < 0.001$). Before using a more complex simulation approach that tries to model the pronunciation process from A to Z, it is therefore useful to see how much explanatory power a simple orthography-to-lexeme network can provide for the current data.

While we decided not to train a network mapping acoustic features onto lexemes, we did use a different type of additional network in the current simulations. In Chapter 3, we found that a lexeme-to-lexeme

network provided explanatory value over and above an orthography-to-lexeme network for the eye movement patterns on compounds in natural discourse reading. As in Chapter 3 we therefore trained a lexeme-to-lexeme discrimination learning network to gauge contextual learning at the lexeme level.

As in Chapter 3, both the orthography-to-lexeme and lexeme-to-lexeme networks were trained on the British National Corpus (henceforth BNC; Burnard, 1995). For the orthography-to-lexeme network the input cues were letter trigrams and the outcomes were lexemes. For the lexeme-to-lexeme network, the input cues were lexemes n-2 and n-1 in a word trigram and the outcome was lexeme n.

We extracted three systemic measures of language processing from the orthography-to-lexeme network. These three measures are the activation of (1) the preposition, (2) the definite article and (3) the target noun given the presentation of the preposition, the definite article and the target noun. We obtained these activations for all of the 272 phrases that were used in the experiment by summing the associations between all letter trigrams in the input phrase and the preposition, the definite article and the target noun lexeme (see Equation 1.3). For the example phrase "into the onion", for instance, we calculated the activation of the target noun "onion" by summing the associations between the letter trigrams *#in*, *int*, *nto*, *to#*, *o#t*, *#th*, *the*, *he#*, *e#o*, *#on*, *oni*, *nio*, *ion* and *on#* (hash marks indicate word boundaries) and the lexeme *ONION*. Similarly, the simulated activations of the preposition "into" and the definite article "the" were defined as the summed association between these letter trigrams and the lexemes *INTO* and *THE*, respectively.

The simulated activations for the preposition, determiner and target noun will henceforth be referred to as *NDL Activation Preposition*, *NDL Activation Determiner* and *NDL Activation Word*. Following Baayen et al. (2011), we applied an inverse and logarithmic transformation to all activations prior to analysis to remove a rightward skew from the data. As such, the activation measures are proportional to the system complexity of the relation between form and meaning. Furthermore, we

added a back off constant of 0.05 to all activations to prevent division by zero when applying the inverse transformation.

As in Chapter 3, we derived a more general systemic property of the target word lexeme from the lexeme-to-lexeme network: the median absolute deviation (henceforth MAD) of the vector of target noun weights given all word types in the training lexicon. We successfully applied the MAD measure in the context of discrimination learning in Chapter 3 as a measure of network connectivity: the greater the MAD of a lexeme, the greater its network connectivity and the easier it is to access that lexeme. As such, one could think of the MAD measure as a systemically motivated account of frequency effects. The greater the frequency of a lexeme, the better a discrimination learning network is able to learn which lexemes are positively or negatively associated with that lexeme (and therefore the greater the MAD). We will henceforth refer to the MAD measure as *NDL MAD*. We log-transformed *NDL MAD* prior to analysis to remove a rightward skew from the *NDL MAD* distribution.

As for the lexical predictor analysis, we removed predictor outliers further than two standard deviations from the mean from the data prior to analysis. As a consequence, we excluded 1.54% of predictor values for *NDL Activation Word*, 5.00% of all predictor values for *NDL Activation Determiner*, 6.92% of all predictor values for *NDL Activation Preposition* and 4.62% of all predictor values for *NDL MAD*. Table 4.2 shows the range, adjusted range, mean, median and standard deviation for all NDL predictors.

As for the lexical predictor data set, the NDL predictors are characterized by a considerable amount of collinearity ($\kappa = 59.97$). Most notably, there is a medium correlation between *NDL MAD* and *NDL Activation Word* ($r = 0.52$). Nonetheless, suppression is unlikely given this correlation. As for the lexical predictor analysis, we therefore decided not to decorrelate the NDL predictors. As for the lexical predictor analysis, we ensured that the effects reported below are robust through a post-hoc analysis. For each NDL predictor, we fitted a new model with an identical model structure, but omitting the other NDL predictors at the electrode

Table 4.2. Summary of the independent variables (*log*) *Picture Complexity*, (log and inverse transformed) *NDL Activation Preposition*, (log and inverse transformed) *NDL Activation Determiner*, (log and inverse transformed) *NDL Activation Word* and (*log*) *NDL MAD*. Range is the original range of the predictors. Adjusted range is the range after removing predictor outliers. Mean, median and sd are the means, medians and standard deviations after outlier removal.

predictor	range	adjusted range	mean	median	sd
Picture Complexity	8.53 - 11.13	8.69 - 10.83	9.88	9.91	0.50
NDL Act. Preposition	−0.65 - −1.70	−0.19 - 0.58	0.04	0.00	0.13
NDL Act. Determiner	−0.20 - 0.17	−0.12 - 0.04	−0.04	−0.04	0.02
NDL Act. Word	0.00 - 2.88	0.13 - 2.88	1.62	1.83	0.76
NDL MAD	−15.12 - −8.88	−14.57 - −9.56	−12.07	−12.15	1.26

visualized below. The results of this post-hoc analysis were qualitatively similar to the effects reported below.

Analogous to the analysis for the lexical predictors, we fitted a GAMM with a main effect smooth for time, by-participant factor smooths for trial and time, and random intercepts for preposition, noun and prepositional phrase to the ERP signal at each electrode. In addition, we included a main effect smooth as well as a tensor product interaction between time and predictor (modeled through `ti()` terms) for each of the predictors *Picture Complexity*, *NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD*. As before, non-linearities in the predictor dimension were limited to 5 knots, non-linearities in the time dimension were limited to 20 knots and the autocorrelation parameter ρ was set to 0.75.

4.9 NDL Simulation Results

In this section, we present the results for the predictors *NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD*. The effect of *Picture Complexity* was highly similar to that reported in the lexical predictor analysis and is therefore not repeated below.

4.9.1 NDL Activation Preposition

Figure 4.15 shows the contour plot of the tensor surface for *NDL Activation Preposition* at electrode *P3*. The partial effect of *NDL Activation Preposition* is characterized by a positivity for prepositions with high (log and inverse transformed) activation values in the first 200 ms after picture onset, which is followed by a negativity for the same prepositions. This effect is highly significant across the left hemisphere, and shows peak amplitudes at left and central parietal and occipital electrodes.

Given that log and inverse transformed *NDL* activations are proportional to naming latencies, whereas frequency measures are inversely proportional to naming latencies, the effect of *NDL Activation Preposition* is qualitatively and topographically similar to the effect of *Preposition*

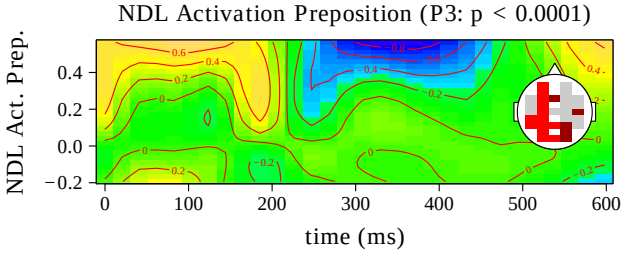

Figure 4.15. Effect for the tensor product interaction between time and (log and inverse transformed) *NDL Activation Preposition* at electrode *P3*.

Frequency described for the lexical predictor analysis. Both predictors show positivities in the first 200 ms followed by negativities at later points in time for predictor values for which longer naming latencies are expected (i.e., high (inverse-transformed) activation, low frequency). In both cases, the effect is present across the left hemisphere, but is most prominent in left-central parietal-occipital areas. The similarity of the effects for *Preposition Frequency* and *NDL Activation Preposition* is unsurprising given the correlation between both predictors ($r = -0.60$).

Consistent with the absence of a main effect of *Preposition Frequency*, we found little evidence for a main effect smooth for *NDL Activation Preposition* in the left hemisphere. In contrast to the main effect of *Preposition Frequency*, however, the main effect of *NDL Activation Preposition*

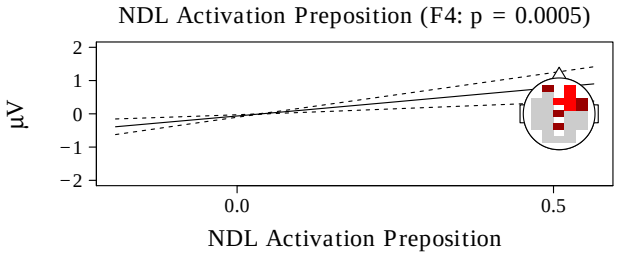

Figure 4.16. Effect for the main effect smooth of (log and inverse transformed) *NDL Activation Preposition* over time at electrode *P3*.

did reach significance at some frontal and frontal-central electrodes in the right hemisphere. Figure 4.16 shows the main effect of *NDL Activation Preposition* over time at electrode *F4*, with more positive voltages for predictor values that correspond to expected processing difficulties (i.e., longer naming latencies).

4.9.2 NDL Activation Determiner

Figure 4.17 presents the time by predictor tensor product interaction for *NDL Activation Determiner*. This effect is characterized by a complicated pattern of oscillatory activity in both the time and predictor dimensions. For a substantial number of time values, the effect is mirrored with respect to the middle of the *NDL Activation Determiner* range. We see a concave effect in the predictor dimension that starts around 80 ms after picture onset and that returns from 220 to 300 milliseconds. After that, the effect reverses, with a convex effect of *NDL Activation Determiner* from 320 ms onwards. This effect is most prominent in left and central parietal-occipital areas, but reaches significance across the left hemisphere.

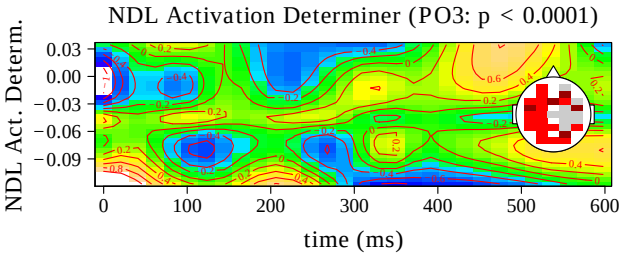

Figure 4.17. Effect for the tensor product interaction between time and (log and inverse transformed) *NDL Activation Determiner* at electrode *PO3*.

Given the fact that the determiner is identical in all stimuli, the presence of a statistically robust time by *NDL Activation Determiner* tensor product interaction with a relatively large effect size may seem surprising from a traditional point of view. From a discrimination learning perspective, however, the effect of *NDL Activation Determiner* is

expected. The *NDL Activation Determiner* measure used here is defined as the activation of the determiner lexeme given the orthographic cues of not only the determiner, but also of the preposition that precedes it and the target noun that follows it. The current effect therefore demonstrates that the context in which a determiner appears has considerable influence on how it is processed.

We furthermore found some evidence for a main effect of *NDL Activation Determiner*. As can be seen in Figure 4.18, voltages tend to be somewhat higher for higher values of *NDL Activation Determiner* at left and central parietal-occipital electrodes. This effect, however, reach significance at a Bonferroni corrected alpha level at 1 electrode only. It is therefore unclear how statistically robust the main effect of *NDL Activation Determiner* is.

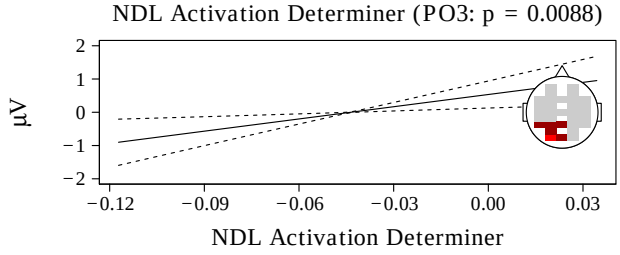

Figure 4.18. Effect for the main effect smooth of (log and inverse transformed) *NDL Activation Determiner* over time at electrode *PO3*.

4.9.3 NDL Activation Word

The time by predictor tensor product interaction for *NDL Activation Word* at example electrode *FC1* is presented in Figure 4.19. The effect is characterized by oscillations at the lower end of the theta range (3 *Hz*). The oscillations are most prominent for high predictor values, but are also present for lower predictor values. The effect of *NDL Activation Word* is topographically widespread, with significant time by *NDL Activation Word* tensor product interactions across the scalp. Peak amplitudes, however, are reached in frontal and central areas in the left hemisphere

and parietal and occipital areas in the right hemisphere. The effect of *NDL Activation Word* first reaches significance at 149 ms after picture onset for medium values of *NDL Activation Word*.

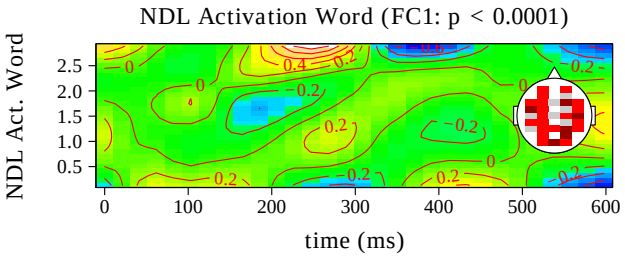

Figure 4.19. Effect for the tensor product interaction between time and (log and inverse transformed) *NDL Activation Word* at electrode *FC1*.

The time by *NDL Activation Word* interaction shows some similarities with the time by *Word Frequency* interaction described earlier. The effect is topographically widespread and characterized by oscillations in the lower part of the theta range. The onset of the effect, however, is later than that of the Word Frequency effect, which was first significant at 97 ms after picture onset. Furthermore, *NDL Activation Word* shows clear non-linearities in the predictor dimension. By contrast, the time by *Word Frequency* interaction was mostly characterized by simple linear effects in the predictor dimension with alternating positive and negatives slopes. The moderate similarities between the effect of *NDL Activation Word* and the effect of *Word Frequency* are in line with the moderate correlation between both predictors (r = -0.41).

For *Word Frequency* we found little evidence for a main effect over time. For *NDL Activation Word*, we found a main effect at 6 electrodes located in bilateral frontal areas only (see Figure 4.20), with somewhat more positive voltages for high predictor values (i.e., for words with longer expected naming latencies). The main effect of *NDL Activation Word* did not reach significance at a Bonferroni-corrected alpha level at any electrodes. In addition, the electrodes at which we saw significant effects at a non-corrected alpha level were limited to frontal electrodes.

Given the increased RMS values at these electrodes these effects need to be interpreted with care. As such, we conclude that the evidence for a main effect of *NDL Activation Word* is limited at best.

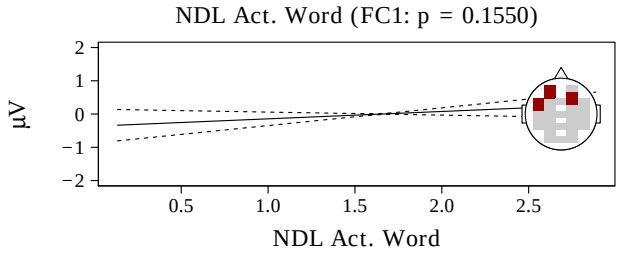

Figure 4.20. Effect for the main effect smooth of (log and inverse transformed) *NDL Activation Word* over time at electrode *FC1*.

4.9.4 NDL MAD

Whereas *NDL Activation Word* describes the bottom-up support for the target noun lexeme, *NDL MAD* is top-down measure of the network connectivity of a word that is perhaps best perceived of as a systemic alternative to word frequency measures that captures the out-of-context probability of a word. Indeed, *NDL MAD* correlates much more strongly with *Word Frequency* ($r = 0.90$) than *NDL Activation Word* ($r = -0.41$). As such, we would expect the effect of *NDL MAD* to be more similar to the effect of *Word Frequency* than the effect of *NDL Activation Word*. As can be seen in Figure 4.21, this prediction is borne out.

The effect of *NDL MAD* is characterized by 3 to 4 *Hz* oscillations for both high and low predictor values. For high values of *NDL MAD* the phase of these oscillations is highly similar to the phase of the oscillations observed for *Word Frequency*. For low predictor values, there is a phase mismatch with the oscillations in the first 250 ms. From 250 to 600 ms after picture onset, however, the phase of the oscillations is highly similar to that of the oscillations for *Word Frequency* once more.

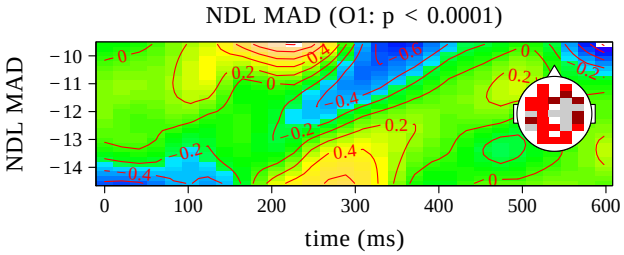

Figure 4.21. Effect for the tensor product interaction between time and (*log*) *NDL MAD* at electrode *O1*.

The topographical distribution of the time by *NDL MAD* interaction is similar to that of the time by *Word Frequency* interaction as well, with a widespread effect that is significant across the left hemisphere, as well as in central and right parietal-occipital areas. Furthermore, the effect of *NDL MAD* at high predictor values is first significant at 100 ms after picture onset. As such, the temporal onset of the *NDL MAD* effect is highly similar to that of the *Word Frequency* effect, which was first significant at 97 ms after picture onset. In conclusion, therefore, the effects of *NDL MAD* and *Word Frequency* show remarkable similarities.

For low values of *NDL MAD*, we see an early negativity that is first significant at 20 ms after picture onset. Perhaps, this effect is an artifact due to the unreliability of GAMMs near the edges of the analysis window. As such, a negativity for low values of *NDL MAD* around 100 ms after picture onset may incorrectly be present in the first 50 ms of the analysis window as well. The *NDL MAD* measure, however, taps into the a priori probability of a word. The early effect of *NDL MAD* may therefore well reflect anticipatory predictions at or even before picture onset.

Figure 4.22 presents the main effect of *NDL MAD* over time. We found no statistically significant evidence for a main effect of *NDL MAD*, neither at a corrected, nor at an uncorrected alpha level. The effect of *NDL MAD* is therefore best described by a time by *NDL MAD* interaction.

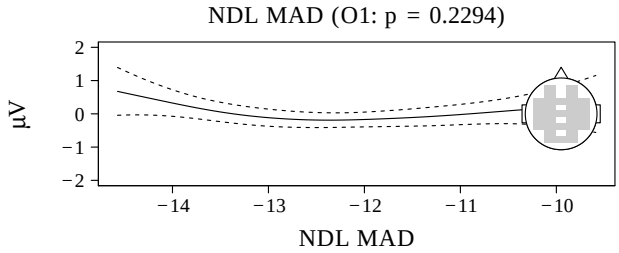

Figure 4.22. Effect for the main effect smooth of (*log*) *NDL MAD* over time at electrode *O1*.

4.10 Discussion

The NDL simulation described above demonstrated that the ERP signature of *Preposition Frequency* as it evolves over time is highly similar to that of *NDL Activation Preposition*, with a qualitatively and topographically similar effects for both predictors. The effects of *Word Frequency* and *NDL Activation Word* showed some similarities as well, but the NDL measure that most closely resembled the pattern of results for *Word Frequency* was *NDL MAD*, a systemic measure of the prior probability of a word. The time by predictor tensor product interactions for both *Word Frequency* and *NDL MAD* showed theta range oscillations with similar phases and similar topographical distributions. The prepositions used in the primes were selected at the quantiles of the prepositional phrase distribution for each target noun. The primes therefore provide little information about the identity of the upcoming target noun. As such, the fact that the *Word Frequency* effect is to a considerable degree driven by the out-of-context probability of the target noun is less than surprising.

The effect of *Phrase Frequency* was characterized by a prolonged effect over time, with higher voltages for lower frequency phrases. The fact that we found some evidence for main effects over time for *NDL Activation Preposition* and – to a lesser extent – *NDL Activation Determiner* may indicate that these systemic measures pick up part of the

prolonged effect observed for *Phrase Frequency*. The effects of both *NDL Activation Preposition* and *NDL Activation Determiner* were inhibitory in nature. Given the inverse transform applied to the activation measures, voltages were therefore lower for prepositions and determiners with more bottom-up support. As such, the qualitative nature of the main effects of *NDL Activation Preposition* and *NDL Activation Determiner* is in line with the fact that high frequency phrases give rise to more negative voltages.[2] The effect of *Phrase Frequency* was topographically widespread, however, whereas the main effects of *NDL Activation Preposition* and *NDL Activation Determiner* had much narrower topographies.

Baayen et al. (2013) simulated the phrase frequency effect in lexical decision in the NDL framework through a simple additive integration of the activations of the component words given the orthographic features in a phrase. A post-hoc analysis at electrode $O1$, however, revealed that the effect of a similar additive integration of *NDL Activation Preposition*, *NDL Activation Determiner* and *NDL Activation Word* did not show a similar prolonged linear main effect of *Phrase Frequency*. While systemic measures of the bottom-up support for the preposition and the determiner may pick up part of the effect, therefore, we conclude that the current NDL approach is unable to account for the (full) effect of *Phrase Frequency* documented here.

In Chapter 3, we similarly found that NDL measures derived from orthography-to-lexeme and lexeme-to-lexeme networks were unable to account for the effect of trigram frequency on the eye fixation patterns during noun-noun compound reading. There, we suggested that the inability to account for n-gram frequency effects may have been a consequence

[2] Note that some variation with respect to the reported main effects exists. At electrode *F8*, for instance, the main effect of *Phrase Frequency* shows the opposite pattern of results as compared to the effect reported for example electrode *O1*. For all main effects, we selected example electrodes that give a good impression of the overall nature of the effect. While the main effect of *Phrase Frequency* at electrode *F8* is qualitatively different from the main effect of *Phrase Frequency* at the reported example electrode *O1*, for instance, the other electrodes that show a significant main effect of Phrase Frequency over time (*Fp1*, *F3*, *T7*, *C3*, *P7*, *P3*, *PO3*, *Oz*) show an effect that is qualitatively similar to the reported effect at electrode *O1*.

of the limited frequency of compound-final trigram in the BNC. To some extent, the same problem may underlie the failure of the NDL measures to account for the phrase frequency observed here. The average frequency of the prepositional phrases in the BNC was 16.65, and the median was no more than 3. No less than 52 of the 272 prepositional phrases used in this experiment (19.12%) never occurred in the BNC and only 15 phrases had a BNC frequency greater than 50. As in Chapter 3, therefore, the size of the BNC may simply be too small to allow discrimination learning networks to pick up on the distributional patterns that underlie phrase frequency effects.

Furthermore, the NDL activation measures capture the orthographic bottom-up support for the lexemes in a prepositional phrase. The current task, however, is not purely orthographic in nature. While the preposition plus definite article primes are presented orthographically, the target noun is presented visually through a photograph of the object that the noun refers to. The phrase frequency effect observed here, therefore, is likely to partly reflect processes related to the uptake of visual information and the integration of this information with the preceding orthographic input. Given that these processes are outside the scope of the lexical learning approach used here, it is unsurprising that the systemic NDL measures cannot (fully) explain the phrase frequency effect reported here.

4.11 Quantitative performance lexical predictors and NDL measures

A final point of interest regarding the NDL simulation reported above is the quantitative performance of the NDL measures as compared to the lexical predictors *Preposition Frequency*, *Word Frequency*, *Relative Entropy* and *Phrase Frequency*. Given the different size of the data sets for both analyses after outlier removal, a direct comparison of the quantitative performance of the models reported above through goodness-of-fit measures was not possible. For the lexical predictor models, we therefore constructed baseline models for the same data set as the ori-

ginal lexical predictor models that had an identical model structure, but that excluded the lexical predictors of interest (*Preposition Frequency*, *Word Frequency*, *Relative Entropy* and *Phrase Frequency*). Similarly, we constructed baseline models for the NDL models that excluded the NDL predictors of interest (*NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD*). We then looked at the difference in deviance explained between the lexical predictor models and the baseline lexical predictor models, as well as between the NDL models and the baseline NDL models.

Generally speaking, the contribution of both the lexical variables and the NDL measures to the deviance explained by the models was small, with improvements in the overall percentage of deviance explained (i.e., deviance explained by full model minus deviance explained by baseline model) being substantially smaller than 1%. The average additional percentage of deviance explained across all electrodes was highly similar for the lexical predictor models (0.100%) and the NDL models (0.098%), with a paired t-test on the vectors of additional deviance explained for all electrodes in the lexical predictor and NDL models showing no significant difference ($p = 0.512$). As such, the quantitative performance of the NDL measures in GAMMs seems comparable to that of standard lexical predictors.

To gain further insight into the relative contribution of the lexical predictors and the NDL measures, we furthermore fit gradient boosting machines (GBMs, see J. H. Friedman, 2001, 2002) as implemented in version 2.1.1 of the *gbm* package for R (Ridgeway, 2015) to the ERP signal at each electrode. Much like random forests, GBMs consist of a large number of regression (or classification) trees. Unlike random forests, however, the trees in GBMs are not independent. Instead, tree n is grown on the basis of the residual error of trees 1 to $n - 1$. GBMs allow for missing data through the use of surrogate splits. We could therefore fit GBMs to the full data set without losing data points due to outlier removal.

For each electrode, we fitted a GBM with *Time*, *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency*, *Relative Entropy*, *NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD* as predictors to the ERP signal after picture onset. Each GBM consisted of $2,500$ trees and was fitted with an interaction depth of 5 and – due to the size of the data set – a fairly aggressive learning rate of 0.10. To prevent overfitting, we used 10-fold cross-validation and set the minimum number of observations in a node to $5,000$.

The mean relative influence across all electrodes for *Time* was 48.57%, and the mean relative influence of *Picture Complexity* was 4.97%. The summed mean relative influence for all 4 lexical predictors was 21.13%, whereas that of the NDL measures was 25.33%. A paired t-test on the summed relative influence of the lexical predictors and the NDL predictors at all electrodes was highly significant ($t(31) = -12.255$, $p < 0.001$). As such, the GBM analysis indicates that the NDL measures are better predictors of the ERP signal following picture onset as compared to the lexical predictors.

Figure 4.23 presents the relative influence of the individual predictors. The bottom-up measures of the support for the lexico-semantic information associated with the preposition, the determiner and the noun all showed substantial mean relative influences (*NDL Activation Preposition*: 8.46%, *NDL Activation Determiner*: 7.59%, *NDL Activation Word*: 5.48%). The a priori probability of the noun had a more modest contribution to the GBMs (*NDL MAD*: 3.80%). Among the lexical predictors, *Phrase Frequency* had the largest mean relative influence (7.84%), followed by *Relative Entropy* (6.18%), *Word Frequency* (3.63%) and *Preposition Frequency* (3.49%).

The NDL measures of the bottom-up support for the preposition and target noun lexemes outperform the corresponding frequency measures by a substantial margin. Paired t-tests on the relative influence of *Preposition Frequency* and *NDL Activation Preposition* ($t(31) = -12.363$, $p < 0.001$) and the relative influence of *Word Frequency* and *NDL Activation Word* ($t(31) = -7.802$, $p < 0.001$) were highly significant. The GBM ana-

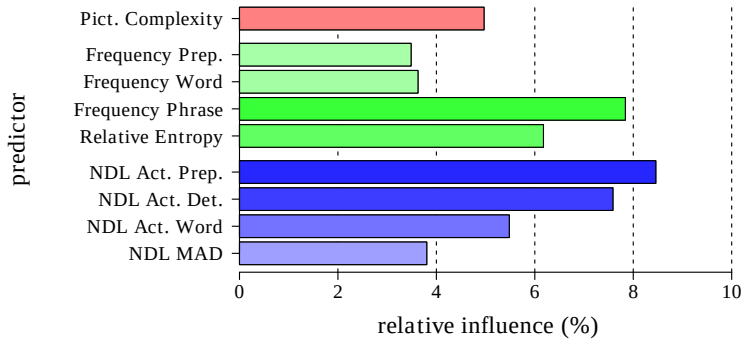

Figure 4.23. Relative influence (%) of *Picture Complexity* (red bar), the lexical predictors *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* (green bars) and the NDL predictors *NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD* (blue bars) in a gradient boosting machine.

lysis therefore suggests that systemic measures of the contextual support for a lexeme are considerably more powerful predictors of the ERP signal following picture onset as compared to simple frequency measures.

4.12 General Discussion

The first half of this chapter discussed the results of a primed picture naming study on prepositional phrase processing. In this experiment participants were presented with preposition plus definite article primes (e.g., "on the") followed by target photographs depicting concrete nouns (e.g., "strawberry"). Participants were asked to name the target noun as fast and accurately as possible. We recorded the ERP signal after picture onset and analyzed the correlates of four linguistic predictors in this signal using generalized additive mixed-effect models.

At the word level we observed significant time by predictor interactions for the frequency of the preposition and the target word, as well as for the prepositional relative entropy of the target word. For word frequency, we observed oscillations in the time dimension with a

frequency at the lower end of the theta range (3-7.5 *Hz*) across the left hemisphere, as well as in bilateral occipital-parietal areas. As mentioned above, theta range oscillations are thought to reflect (working) memory demands in language processing that arise from the synchronous firing of neurons in hippocampal areas (see Bastiaansen & Hagoort, 2003) and have previously been observed in a variety of language processing tasks (see, e.g., Bastiaansen et al., 2005, 2008; Grabner et al., 2007). The effect of target word frequency was first significant at 97 ms after picture onset. This early onset of the word frequency effect is in line with previous studies that established the onset of word frequency effects (Hauk et al., 2006; Penolazzi et al., 2007; Sereno et al., 1998) soon after the 100 ms mark.

Of the word level effects, the effect of relative entropy is of particular theoretical interest. Previously, relative entropy effects had only been observed in reaction time studies (see, e.g., Milin, Filipović Durdević & Moscoso del Prado Martín, 2009; Milin, Kuperman et al., 2009; Kuperman et al., 2010; Baayen et al., 2011). The current study is the first to document a relative entropy effect in an ERP study, with oscillations near the lower edge of the theta range that were most prominent in parietal and occipital areas. These oscillations had greater amplitudes for high predictor values as compared to low predictor values. Similar to the reaction time studies mentioned above, therefore, the current results suggest that additional processing is necessary when a noun's use of prepositions is less prototypical. The effect of relative entropy emerged early, showing a significant effect as early as 95 ms after picture onset. The temporal onset of the relative entropy effect is therefore similar to that of word frequency (97 ms after picture onset).

At the phrase level, we observed an effect of phrase frequency that was qualitatively different from the effect of word frequency. While the word frequency effect was characterized by oscillations in the time domain, the phrase frequency effect is best described as a near-linear prolonged effect over time with more positive voltages for low frequency phrases and more negative voltages for high frequency phrases. This effect was most prominent in left-lateralized parietal and occipital areas.

As for the effect of relative entropy, the effect of phrase frequency is well-documented in chronometric studies (see e.g., Bannard & Matthews, 2008; Arnon & Snider, 2010; Siyanova-Chanturia et al., 2011; Tremblay et al., 2011; Shaoul, Westbury & Baayen, 2013). Recently, Tremblay and Baayen (2010) documented a phrase frequency effect in an ERP study for 4-word sequences using a free recall task. The current study adds to these findings with a phrase frequency effect in a primed picture naming paradigm.

Effects of n-gram frequency provide evidence for "some experience-derived knowledge of specific [...] word sequences" (Bannard & Matthews, 2008, p.246). How this knowledge is implemented, however, is not clear. One possibility is that phrase representations are stored holistically, much like word representations. As noted by Baayen et al. (2013), such a perspective on n-gram frequency effects fits well with the architecture of exemplar-based approaches to language processing, such as data-oriented parsing (Bod, 2006) or memory-based learning (Daelemans & Bosch, 2005). Although the current results are not incompatible with exemplar-based models, these models would have to assume storage of (and computation over) many millions of n-gram representations to account for n-gram frequency and relative entropy effects (see Baayen et al., 2013). Furthermore, if word representations and phrase representations are stored and accessed in the same way we would expect the effects of word frequency and phrase frequency to be highly similar. Here, however, we found qualitatively different results for word and phrase frequency.

Discrimination learning offers an alternative interpretation of phrase frequency and relative entropy effects. In the Naive Discriminative Reader NDR model (Baayen et al., 2011) no representations beyond the simple word level exist. Nonetheless, the NDR correctly predicts the chronometric effect of phrase frequency (Baayen et al., 2013). The second part of this chapter presents a simulation study in which we investigate the explanatory power of the NDL framework for the current data. We constructed statistical models similar to those for the lexical predictor analysis. The lexical predictors, however, were replaced by 4 measures derived from discrimination learning networks: 1 measure regarding the out-of-context

probability of the target word, and 3 measures gauging the amount of bottom-up support for the preposition, the determiner and the target noun given the presence of all three words in the visual input.

The bottom-up support for the preposition showed an effect that was qualitatively and topographically similar to the effect of *Preposition Frequency*. We also observed some similarities between the effect for the bottom-up support for the noun and the effect of *Word Frequency*. The systemic measure that most closely resembled the effect, however, was a top-down measure of the out-of-context probability of the target noun. As indicated by an analysis using gradient boosting machines (GBMs), the measures of the bottom-up support for the preposition and the noun had increased explanatory power as compared to *Preposition Frequency* and *Word Frequency*. Taking the systemic support for a word in its linguistic context into account, therefore, helps better understand the linguistic processes that underlie the ERP signal after picture onset.

Phrase frequency does not map 1-to-1 with any of the discrimination learning measures used here. Nonetheless, there is some evidence that at least part of the phrase frequency effect is captured by the discrimination measures used here. The predicted values of a GAMM with *Phrase Frequency* as the dependent variable and simple main effect smooths ($k = 5$) of the four systemic measures described above as independent variables show a moderate correlation of $r = 0.56$ with *Phrase Frequency*. Furthermore, we found some evidence for main effect smooths of the bottom-up support for the preposition and – to a lesser extent – the determiner that were qualitatively similar to the prolonged effect of *Phrase Frequency*.

It is clear, however, that the systemic measures used here do not capture the full effect of *Phrase Frequency* in the current study. The topographical distribution of the above-mentioned main effects of the bottom-up support for the preposition and the determiner are much more confined than that of *Phrase Frequency*. Furthermore, while Baayen et al. (2013) showed that an additive integration of the bottom-up support for all words in a phrase captured the qualitative nature of the phrase frequency effect in lexical decision, a similar additive integration of the bottom-up support for the preposition, the determiner and the noun

did not yield a prolonged effect similar to the *Phrase Frequency* effect reported here. These shortcomings of the NDL measures with respect to the phrase frequency effect may originate from the limited size of the training data for the NDL networks and/or the cross-modal experimental paradigm adopted here.

A quantitive comparison of the lexical distributional variables and the NDL measures indicated that the contribution of both sets of predictors in GAMMs was highly similar. By contrast, a series of GBMs fit to each electrode indicated that the NDL measures co-determined the ERP signal following picture onset to a greater extent than did the lexical predictors. We therefore conclude that discrimination learning offers a competitive framework for understanding the ERP signal in the primed picture naming paradigm. It is important to note, however, that the NDL framework – much like an analysis based on lexical distributional variables – provides a high-level functional window on lexical processing that tells us little about the neuro-biological implementation of the discriminative learning mechanism it posits. The discrete representations in the NDL framework do not do justice to the complex architectural and topographical neuro-biological reality of neural networks. Nonetheless, the current simulations demonstrate that systemic measures derived from discrimination learning networks provide further insight into the behavioral effects of lexical predictors and advance our understanding of the language processing system. When trying to understand the complex dynamic system that language is, there is no harm in starting with the basic principles of learning.

5

Conclusions

Chapter 1 of this dissertation introduced the naive discrimination learning (NDL) framework as a symbolic approach to investigating the role of learning in language processing and provided a brief overview of other psycholinguistic studies that have investigated language processing from a discrimination learning perspective, either prior to or in parallel with the work presented here. The following three chapters consisted of applications of the NDL approach to three psycholinguistic data sets, each involving a different measure of language processing and a different experimental task.

Chapter 2 explored to what extent an NDL model could provide further insight into naming latencies in the reading aloud task. Leading models of reading aloud, including the CDP+ model, are dual-route models (see, e.g., M. Coltheart et al., 2001; Perry et al., 2007, 2010; Harm & Seidenberg, 2004). Words are read through a lexical architecture, in which the connection between orthographic and phonological units is mediated by lexical representations. Non-words - or words unknown to a reader - by contrast, are read through a non-lexical route in which orthographic features are mapped directly onto phonological features.

To simulate reading aloud in the NDL framework, I extended the NDR model of Baayen et al. (2011) for silent reading with a lexeme-to-phonology network. In contrast to dual-route models, the resulting NDR_a model consists of a single lexical route that is responsible for both word and non-word naming. Whenever a word is presented, the orthographic units in the input activate the lexical representation of the target word, as well as the lexical representations of orthographic neighbors. These lexical representations then activate phonological units, which allow for the pronunciation of the word. For non-words or for words that were not previously encountered by the reader, no lexical representation of the target word exists. The pronunciation of these words therefore relies entirely on the phonological units activated by the lexical representations of orthographic neighbors of the target word.

The single-route NDR_a model showed similar correlations with the observed naming latencies as the leading dual-route model of reading aloud, the CDP+ model (Perry et al., 2007), and accurately captured a number of effects related to the consistency of the orthography to phonology mapping that were previously interpreted as evidence for the existence of a non-lexical route (see Perry et al., 2007). Furthermore, the NDR_a model correctly predicts a hitherto unobserved effect of non-word frequency. As such, the NDR_a model provides a highly competitive single-route alternative to existing dual route models of reading aloud.

Chapter 2 furthermore demonstrates that the addition of a sub-lexical route does not help further improve the performance of the NDR_a model. This finding stands in contrast to Perry et al. (2007), who found that the sub-lexical route of the CDP+ model had a substantial contribution to the performance of the model. Perry et al. (2007) implemented a learning network in the sub-lexical route of the CDP+ model (see Zorzi et al., 1998b, 1998a). The lexical route of the model, however, is based on the interactive activation model of McClelland and Rumelhart (1981) and therefore does not account for learning. The fact that a sub-lexical route did not improve the performance of the NDR_a model suggests that the contribution of the sub-lexical network of the CDP+ model may not be evidence for the psychological reality of a sub-lexical route, but instead

for the shortcomings of a less-than-optimal implementation of the lexical route. As such, the findings discussed in Chapter 2 demonstrate that investigating language processing from a systemic perspective may lead to interesting insights into the nature of language processing that could not be obtained through more traditional approaches.

Chapter 3 presented the results of an analysis of the eye movement patterns during compound reading in natural discourse. Systemic measures of the bottom-up support for noun-noun compounds and their a priori probability were extracted from two discrimination learning networks: an orthography-to-lexeme network and a lexeme-to-lexeme network. Across different measures of the eye movement patterns, including the duration and position of first, second and third fixations, these systemic measures provided explanatory power that was comparable to that of an extensive set of lexical predictors.

Fixation patterns during first fixations were primarily determined by the bottom-up support for the relevant lexico-semantic information given the orthographic features that were available to the reader. The duration of first-and-only fixations was co-determined by a weighted sum of the activation of the modifier, the head and the compound lexeme given the orthographic features of the compound as a whole. Compared to first-and-only fixations, first-of-many fixations were characterized by more leftward fixation positions and larger incoming saccade sizes. As a result of the suboptimal viewing position, not all orthographic features were available to the reader. First-of-many fixations were therefore influenced by the activation of the modifier lexeme given the first orthographic trigram of the compound only.

The pattern of results for first fixation durations demonstrates that fixation patterns are to a large extent determined by the amount of information that is available in the spotlight of visual attention. The NDL framework offers the opportunity to explicitly specify which orthographic features are available to the readers. This allows for a more precise understanding of first fixation patterns as compared to standard lexical predictors. Nonetheless, the all-or-none availability of input features is

an oversimplification of what is actually available to the eye (I will return to this issue shortly).

In a standard lexical predictor analysis, both first-and-only and second fixation durations showed an effect of compound frequency on fixation duration. The NDL analysis, however, suggested that these frequency effects are qualitatively different. Whereas first-and-only fixations durations showed an effect of the bottom-up support for the lexico-semantic information associated with the compound and its constituents, second fixation durations were co-determined by the a priori probability of the compound. As such, readers seem to fall back on a top-down "best guess" strategy when compound processing is suboptimal. The systemic measures derived from NDL networks, therefore, provided a more precise and more differentiated account of the eye fixation data as compared to standard lexical predictors.

Chapter 4 discussed the analysis of the ERP signal following picture onset in a primed picture naming task. Participants were presented with preposition plus definite article primes (e.g., "in the") followed by photographs depicting concrete nouns (e.g., "strawberry") and were asked to name the target noun as quickly and accurately as possible. A standard lexical predictor analysis of the data revealed an effect of preposition frequency, as well as oscillatory activity tied to the frequency of the target word and the prototypicality of the frequency distribution of the prepositional paradigm of the target word. By contrast, the frequency of the prepositional phrase as a whole showed a prolonged effect over time, with lower voltages for high frequency phrases.

In an NDL analysis of the picture naming data, measures of the bottom-up support for the preposition and the target word outperformed frequency measures of these words. The effect of the bottom-up support for the preposition was qualitatively highly similar to the effect of preposition frequency. Furthermore, the effect of the bottom-up support for the target noun showed qualitative resemblances to the effect of word frequency. The systemic measure that yielded a pattern of results most similar to word frequency, however, was a top-down measure of the prior probability of the compound. As in Chapter 3, therefore, the NDL analysis

shed further light on the systemic source(s) of the word frequency effect. Overall, the explanatory power of the NDL measures was comparable to that of the lexical predictors in generalized additive models (GAMs) and significantly better than that of lexical predictors in gradient boosting machines (GBMs).

Taken together, the work presented here shows that a discrimination learning approach is highly competitive with more traditional perspectives on adult language processing, across a variety of experimental paradigms and response variables. Furthermore, the NDL analyses of psycholinguistic data sets presented throughout this dissertation uncover information about the nature of the language processing system that is not available through traditional analysis techniques. For the data sets presented here, this information affords reconceptualizations of language processing that are more intuitive and simpler than existing theories.

The discrimination learning approach used here, however, is in many ways quite far removed from the Platonic ideal of a systemic model of language processing. For one, in the simulations reported here the temporal dimension of the information uptake process was largely ignored. Typically, we assumed that all information in the visual input was simultaneously available. Language processing, however, is critically dependent on the information uptake process. By definition, therefore, bottom-up information becomes available (and fades out of attention) in a sequential fashion.

Chapter 3 addressed the temporal dimension of the information uptake process to some extent. For first-of-many fixation durations on noun-noun compounds, we found that the activation of the modifier lexeme given the first orthographic trigram only provided maximum explanatory power. This demonstrates that a closer inspection of the bottom-up information that is available to a language user at a given point in time can result in improved performance of discrimination learning measures. Nonetheless, all-or-none availability of bottom-up information is a clear oversimplification of the dynamic information uptake process. In reality, at a given point in time, each piece of bottom-up information is available

to a certain extent, ranging from zero availability to full availability on a gradient scale (see, e.g., Engbert et al., 2005).

Second, both Chapters 3 and 4 demonstrated that the type of orthography-to-lexeme and lexeme-to-lexeme NDL networks used in this dissertation have trouble capturing n-gram frequency effects. This may to a large extent be due to the limited size of the British National Corpus (BNC), which was used to train these networks. At 100 million words, the BNC may be too small to provide discrimination learning networks with the opportunity to properly learn the lexical co-occurrence patterns that underlie n-gram frequency effects. Future research may indicate that more training data will allow the NDL framework to accurately capture n-gram frequency effects. Alternatively, it may be the case that additional (measures of) processing mechanisms that integrate systemic measures at the word level are necessary to fully capture the effects of n-gram frequency a discrimination learning framework.

Third, looking at the bottom-up support for a specific lexico-semantic representation provides a narrow window on a much broader and much more complex system. Chapter 3 touched on this subject, showing that an integrative measure of the bottom up support for not only the compound lexeme, but also the modifier and head lexemes proved most predictive for the eye fixation patterns on noun-noun compounds. Even integration over the bottom-up support for multiple lexico-semantic representations, however, involves an obvious simplification. Ultimately, what determines the success of language processing is the extent to which the state of the entire system allows for an adequate understanding of the current – linguistic or non-linguistic – input. Exploring options for gauging the overall state of discrimination learning networks is an interesting challenge for future research.

Baayen et al. (2011) investigated the explanatory power of discrimination learning for the behavior of language users in the lexical decision task. This dissertation consists of an application of the discrimination learning approach to three further experimental tasks: silent reading, reading aloud and picture naming. Together these studies demonstrate that discrimination learning measures provide an insightful and compet-

itive perspective on language processing across a variety of experimental paradigms. Much more research, however, is required to acquire a comprehensive appreciation of the successes and shortcomings of a discrimination learning approach to language processing. I therefore end this dissertation with the words that started it: “Data! Data! Data!”.

Appendices

Appendix A

Comparison of GAMMs and traditional ERP analyses

We used generalized additive models (GAMMs) to analyze the ERP data for the current experiment (Hastie & Tibshirani, 1986; Wood, 2006). Unlike traditional ERP analysis techniques, GAMMs allowed us to investigate the non-linear effects of numerical predictors as they evolve over time. By contrast, traditional ERP analyses typically operate on the basis of dichotomized versions of numerical predictors such as word frequency, phrase frequency or relative entropy. The average curves for the dichotomized predictors values are then compared in by-item or by-subject analyses (i.e., low frequency versus high frequency). In this appendix we compare the performance of GAMMs to the performance of a traditional analysis method for simulated data, as well as for some of the key predictor effects documented in this paper. We demonstrate that the patterns of results for both types of analyses converge in some cases, but that a traditional analysis results in a loss of information and induces dichotomization artifacts in other cases.

First, consider the simulated predictor effect in the top left panel of Figure A.1 The effect is characterized by a two-dimensional sinusoid, with oscillations in both the time and the predictor dimension. White noise with a mean of 0 and a standard deviation of 0.5 was added to each simulated data point. The middle panel of the left column of Figure A.1 shows the results of a GAMM analysis on this simulated predictor effect. The two-dimensional sinusoid in the simulated data is replicated in the GAMM analysis. The frequencies of the oscillations in both directions and the effect sizes match those in the simulated data.

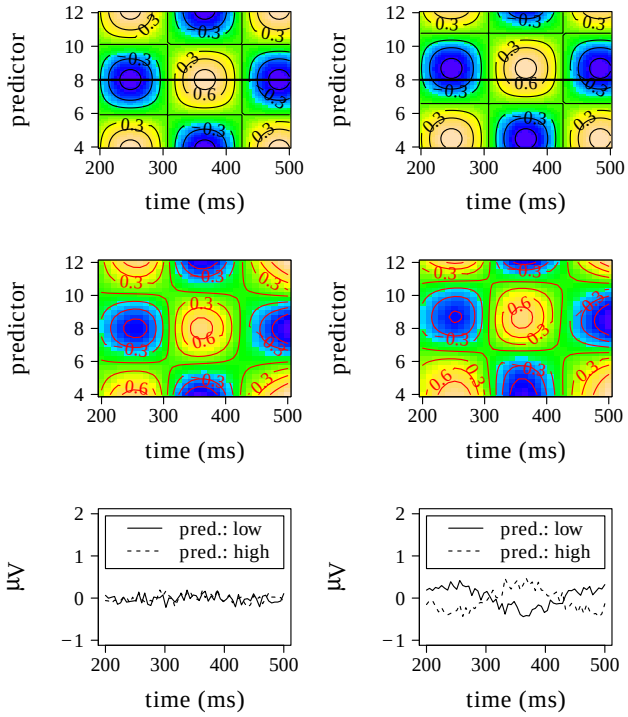

Figure A.1. Simulated predictor effect with an oscillation in both the time and predictor dimension (top panels) and model fits for this effect in a GAMM analysis (middle panels) and a traditional analysis using predictor dichotomization (bottom panels).

The bottom left panel of Figure A.1 shows the results of a dichotomization of the predictor into low and high predictor values based on a split halfway the predictor range. No sinusoidal activity is seen for either high or low frequency words and no difference is observed between high and low frequency words at any point in time. Dichotomization of the predictor therefore entirely masks the two-dimensional oscillatory activity that is present in the simulated data.

The simulated data in the top left panel of Figure A.1 are symmetrical with respect to the mid-point of the predictor range. For the top right panel of Figure A.1 we shifted the effect upwards on the y-axis, such that the simulated predictor effect is no longer symmetrical with respect to the mid-point of the predictor range. The middle panel of the right column of Figure A.1 demonstrates that this does not constitute a problem for GAMMs. As before the two-dimensional sinusoid is replicated with the correct frequency in both dimensions and the correct effect size. The bottom right panel of Figure A.1 shows what happens if the predictor is dichotomized into high and low predictor values with a split at the mid-point of the predictor range. Due to the vertical shift of the oscillations a traditional analysis now reflects some of the oscillatory activity in the simulated data. The observed differences between high and low predictor values, however, reflect the differences between medium and low predictor values in the simulated data. All information about the fact that high predictor values and low predictor values show a highly similar pattern of results is lost.

More subtle examples of the problems associated with the dichotomization of numerical predictors outlined above arise in the ERP data reported here. In what follows, we examine the performance of a traditional ERP analysis for the word frequency, phrase frequency and relative entropy effects in the current data. For each of these three predictors, we compare the GAMM analysis in this chapter to a traditional analysis of the data at the same electrode.

The top panel of Figure A.2 shows the effect of *Word Frequency* at electrode *O1* in the GAMM analysis. The effect is characterized by 3 *Hz* oscillations for both high and low frequency words with opposite phases.

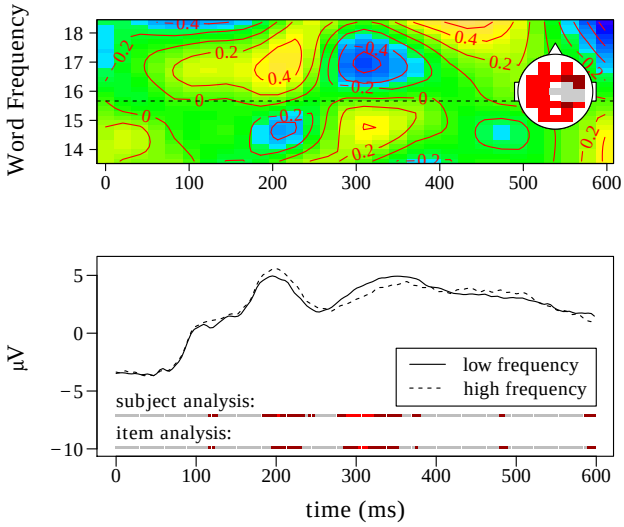

Figure A.2. The effect of *Word Frequency* at electrode *O1* in a GAMM analysis (top panel) and a traditional analysis in which *Word Frequency* is dichotomized (bottom panel). Color coding at the bottom of the second panel indicates significance of the *Word Frequency* effect in item and subject ANOVAs for each point in time.

The dashed line indicates the mean value of *Word Frequency*. The bottom panel of Figure A.2 shows the results of a traditional analysis in which we dichotomized *Word Frequency* into high and low frequency words (split with respect to the mean value of *Word Frequency*). In this analysis we investigated the significance of the dichotomized Word Frequency predictor for each sample point in the time domain by running subject and item ANOVAs on a subset of the data that included all measurements for that sample point, as well as for the previous sample point and the next sample point. The significance of the *Word Frequency* effect in these subject and item analyses is indicated by dark red ($\alpha = 0.05$) and bright red (Bonferroni-corrected alpha level; $\alpha = 0.0016$) in the second panel of Figure A.2.

The grand mean curves for *Word Frequency* show a similar pattern of results as compared to the GAMM analysis. The difference between high and low frequency phrases first reaches significance at a non-corrected alpha level at 117 ms after picture onset, with higher voltages for high frequency phrases. As such, the temporal onset of the Word Frequency effect is somewhat later than the temporal onset of the Word Frequency effect in the GAMM analysis (97 ms after picture onset), presumably due to a loss of statistical power as a result of the predictor dichotomization. Overall, the patterns of results in the GAMM analysis and the dichotomization analysis are highly similar, with higher frequency words showing higher voltages from 180 to 260 ms after picture onset, lower voltages from 260 to 400 ms and higher voltages once more from 400 to about 530 ms (as compared to lower frequency words). Both in the GAMM analysis and in the traditional analysis, the effect of *Word Frequency* is most pronounced from 180 to 400 ms after picture onset.

The comparison of the GAMM analysis and the traditional analysis for the *Word Frequency* effect demonstrates that the oscillatory effect of *Word Frequency* is reflected in the grand means curves for high and low frequency words. Rather than being interpreted as theta range oscillations, however, this effect would likely be described in terms of ERP components in a traditional analysis - with an increased *P200* and a decreased *P350* for high frequency words as compared to low frequency words.

The effect of Word Frequency in the GAMM analysis is relatively simple in nature, with oscillations for high and low frequency words that are nicely separated with respect to the middle of the Word Frequency range and that have opposite phases. This is close to an ideal scenario for a traditional ERP analysis. The effect of *Relative Entropy* represents a somewhat more complicated scenario. The top panel of Figure A.3 shows the effect of *Relative Entropy* at electrode *CP1* in the GAMM analysis.

As can be seen in the top panel of Figure A.3, the effect of *Relative Entropy* is characterized by oscillations in the time domain that arise around 100 ms after picture onset. The oscillations are most prominent for high predictor values, but lower amplitude oscillations are present for

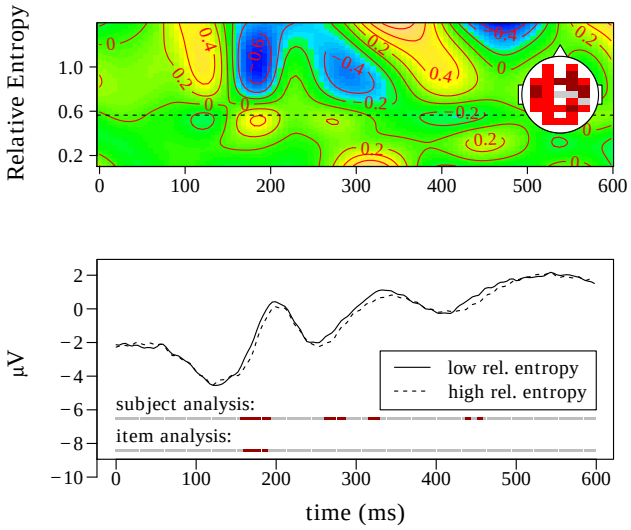

Figure A.3. The effect of *Relative Entropy* at electrode *CP1* in a GAMM analysis (top panel) and a traditional analysis in which *Relative Entropy* is dichotomized (bottom panel).

medium-to-low and low predictor values as well. To complicate things further, the phase difference between the oscillations for high predictor values and the oscillations for low predictor values is not constant, due to small differences in the frequencies of these oscillations.

The bottom panel of Figure A.3 shows the effect of *Relative Entropy* at electrode *CP1* in a traditional ERP analysis in which we dichotomized *Relative Entropy* into high and low relative entropy on the basis of a split at the mean (see dashed line in the top panel of Figure A.3). The grand mean curves for high and low *Relative Entropy* capture the fact that high values of *Relative Entropy* correspond to lower voltages from 150 to 220 ms, from 250 to 340 ms and from 420 to around 500 ms after picture onset (as compared to low values of *Relative Entropy*), although these effects reach significance at non-corrected alpha level only.

The traditional analysis fails to pick up on the more positive voltages for high values of *Relative Entropy* around 100 and 400 ms after picture onset. Potentially, this is due to the fact that only *Relative Entropy* was entered into the traditional analysis, whereas the GAMM analysis uses a multiple regression approach. As such, the effects of other predictors are not taken into account in the traditional analysis. The main effect of phrase frequency, for instance, was marginally significant at electrode *CP1*, $p = 0.077$). Given the nature of the phrase frequency effect (i.e., lower voltages for higher frequency phrases) and the negative correlation between *Relative Entropy* and *Phrase Frequency* ($r = -0.19$), the grand average curve for high values of relative entropy in Figure A.3 may be somewhat lower than it would be if the effect of Phrase Frequency was properly accounted for.

Whereas the qualitative nature of the effect of *Word Frequency* was accurately captured by a traditional ERP analysis, a lot of detail is lost about the effect of *Relative Entropy* through dichotomization. While it might be possible to tell that the *Relative Entropy* effect is characterized by theta range oscillations from the bottom panel of Figure A.3, for instance, it would be impossible to tell that these oscillations are most prominent for high predictor values. Furthermore, the nature of the effect across the predictor dimension is lost through dichotomization. The information that the effect of *Relative Entropy* is U-shaped in nature around 320 ms, for instance, cannot be retrieved from the bottom panel of Figure A.3.

Theta range oscillations in the time dimension characterized the effects of *Word Frequency* and *Relative Entropy*. For *Phrase Frequency*, we found a near-linear effect that persisted over time. The top panel of Figure A.4 shows the additive contour surface of the main effect smooth for *Phrase Frequency* and the time by *Phrase Frequency* tensor product interaction at electrode *O1*, with a long-lasting positivity for low frequency words and a long-lasting negativity for high frequency words.

The bottom panel of Figure A.4 shows the results of a traditional ERP analysis in which *Phrase Frequency* was dichotomized with respect to the mean predictor value (see dashed line in the top panel of Figure A.4).

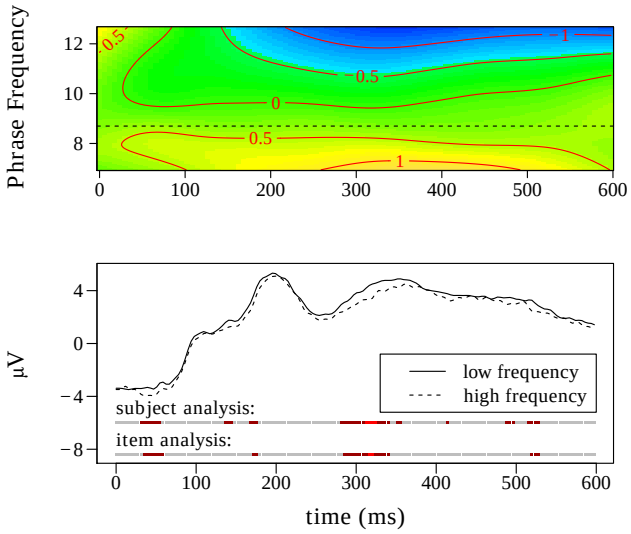

Figure A.4. The effect of *Phrase Frequency* at electrode *O1* in a GAMM analysis (top panel) and a traditional analysis in which *Phrase Frequency* is dichotomized (bottom panel).

The general nature of the *Phrase Frequency* effect is similar to that in the GAMM analysis, with more positive voltages for low frequency words as compared to high frequency words. Consistent with the top panel of Figure A.4, the difference between high and low predictor values is greatest around 300 ms after picture onset, with significant effects in both the item and the subject analysis.

At other points in time, the grand mean curve for high frequency phrases is below that for low frequency phrases as well, but this difference reaches significance for a limited number of sample points at a non-corrected alpha level only. The inability of the subject and item analyses to pick up on the phrase frequency effect throughout the analysis window may be the result of a loss of statistical power in the traditional analysis as compared to the GAMM analysis. This loss in statistical power is a consequence of both the dichotomization of phrase frequency and the fact that other parts of the ERP are not properly controlled for in

the traditional analysis (e.g., trial-effects and random effects of subject, preposition, target noun and prepositional phrase).

In this appendix we compared the GAMM analyses reported in this chapter to traditional ERP analyses using predictor dichotomization for simulated data, as well as for some of the key effects reported in this chapter. Generally speaking, two conclusions can be drawn from this comparison. First, the GAMM analyses reported here seem to provide estimates of predictor effects that are compatible with the grand mean curves. The results of a GAMM analysis and a traditional analysis typically converge as long as dichotomization of a predictor is relatively unproblematic given the nature of a predictor effect. When this is not the case, the differences that arise between the results from a GAMM analysis and a traditional analysis are easily explained given the nature of the predictor effect disclosed by the GAMM.

Second, a GAMM analysis provides much more information as compared to a traditional analysis in which predictors are dichotomized. In a dichotomization analysis, predictor values with very different patterns of results are grouped together, which can result in a loss of statistical power, especially when other sources of variance in the ERP signal are not (properly) taken into account. In addition, the nature of tri- or multipartite predictor effects is - by definition - lost when a predictor is dichotomized. This can lead to a loss of information or misguided conclusions about the nature of an effect. By contrast, as seen in the analysis of the simulated data, GAMM analyses accurately capture non-linear predictor effects as they evolve over time.

Some of the problems associated with a traditional dichotomization analysis can be overcome by choosing an experimental design that investigates the effect of a single categorical predictor with carefully selected predictor values that fall into two or more discrete categories. Many of the questions in psycholinguistic research, however, are easier to answer in multiple regression designs that allow for the simultaneous investigation of the effect of multiple numerical predictors with continuous distributions. The experimental design and analysis techniques presented here provide an example of how the multiple regression techniques that have become

commonplace in reaction time studies can be applied in ERP studies through the use of GAMMs. As demonstrated in this appendix, the results from such a GAMM analysis provide precise information about the linear and non-linear nature of the effects of multiple numerical predictors as they evolve over time.

References

Adelman, J. S. & Brown, G. D. A. (2008). Methods of testing and diagnosing model error: Dual and single route cascaded models of reading aloud. *Journal of Memory and Language*, *59*, 524–544.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Akaike, H. (1980). Likelihood and the Bayes procedure. In J. Bernardo (Ed.), *Bayesian statistics* (pp. 143–166). Valencia: University Press.

Allen, M. & Badecker, W. (2002). Inflectional regularity: Probing the nature of lexical representation in a cross-modal priming task. *Journal of Memory and Language*, *46*, 705–722.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 802–814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 234–254.

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, *35*, 775–800.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, *4*, 439–461.

Andrews, S., Miller, B. & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, *16*(1), 285–311.

Andrews, S. & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1052–1086.

Arnon, I. & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. In *Proceedings of the 31st annual meeting of the cognitive science society* (pp. 2112–2117).

Arnon, I. & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes.* Cambridge, Mass.: The MIT Press.

Baayen, R. H. (2011a). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*, 295–328.

Baayen, R. H. (2011b). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". [Computer software manual]. Available from `http://CRAN.R-project.org/package=languageR` (R package version 1.4)

Baayen, R. H. (2014). Experimental and psycholinguistic approaches to studying derivation. In R. Lieber & P. Stekauer (Eds.), *Handbook of derivational morphology* (pp. 95–117). Oxford: Oxford University Press.

Baayen, R. H., Feldman, L. & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Baayen, R. H., Hendrix, P. & Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech*, *56*(3), 329-347.

Baayen, R. H., Kuperman, V. & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding.* Amsterdam/Philadelphia: Benjamins.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P. & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–482.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX lexical database (cd-rom).* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H. & Schreuder, R. (1999). War and peace: morphemes and full forms in a non-interactive activation parallel dual route model. *Brain and Language*, *68*, 27-32.

Baayen, R. H. & Schreuder, R. (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences)*, *358*, 1-13.

Baayen, R. H., Shaoul, C., Willits, J. & Ramscar, M. (2015). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience*, in press.

Baayen, R. H., Tremblay, A. & Hendrix, P. (2015). An introduction to analyzing the ERP signal with generalized additive modeling using the GAM-eRp package, vignette for the GAM-eRp package for R. *Vignette and package in preparation.*

Balota, D. A. & Chumbley, J. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or pronunciation? *Journal of Memory and Language*, *24*, 89–106.

Balota, D. A., Cortese, M., Sergent-Marshall, S., Spieler, D. & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, *133*, 283–316.

Balota, D. A., Cortese, M. J. & Pilotti. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society* (p. 44). Los Angeles, CA: Psychonomic Society.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. I., Kessler, B., Loftis, B. et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Bannard, C. & Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241-248.

Bastiaansen, M., Berkum, J. v. & Hagoort, P. (2002). Syntactic processing modulates the theta rhythm of the human EEG. *NeuroImage*, *17*(3), 1479-1492.

Bastiaansen, M. & Hagoort, P. (2003). Event-induced theta-responses as a window on the dynamics of memory. *Cortex*, *39*(4-5), 967-992.

Bastiaansen, M., Oostenveld, R., Jensen, O. & Hagoort, P. (2008). I see what you mean: theta power increases are involved in the retrieval of lexical semantic information. *Brain and language*, *106*(1), 15-28.

Bastiaansen, M., Van Der Linden, M., Ter Keurs, M., Dijkstra, T. & Hagoort, P. (2005). Theta responses are involved in lexical-semantic retrieval during language processing. *Journal of Cognitive Neuroscience*, *17*(9), 530-541.

Bertram, R. & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language*, *615-634*, 48.

Binder, J. R., McKiernan, K. A., Parsons, M. E., Westbury, C. F., Possing, E. T. & Kaufman, J. N. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience*, *15*, 372–393.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*, in press.

Borowsky, R., Owen, W. & Masson, M. (2002). Diagnostics of phonological lexical processing: Pseudohomophone naming advantages, disadvantages, and baseword frequency effects. *Memory and Cognition*, *30*, 969–987.

Botvinick, M. M., Cohen, J. D. & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Science*, *8*, 539–546.

Brants, T. & Franz, A. (2006). *Web 1T 5-gram. Version 1.* Philadelphia: Linguistic Data Consortium.

Browman, C. P. & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Browman, C. P. & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*, 201–251.

Browman, C. P. & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*, 299-320.

Browman, C. P. & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155–180.

Brown, G. D. A. (1987). Resolving inconsistency: A computational model of word naming. *Journal of Memory and Language*, *26*, 1–23.

Brunswick, N., McCrory, E., Price, C. J., C.D., F. & U., F. (1999). Explicit and implicit processing of words and pseudowords by adult developmental dyslexics: A search for Wernicke's wortschatz. *Brain*, *122*, 1901–1917.

Burnard, L. (1995). *Users guide for the British National Corpus.* Oxford university computing service: British National Corpus consortium.

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language production (Vol. ii): Development, writing and other language processes* (p. 257-294). London: Academic Press.

Chalmers, D. J. (1992). Subsymbolic computation and the Chinese room. In J. Dinsmore (Ed.), *The symbolic and connectionist paradigms: Closing the gap* (pp. 25–48). London: Lawrence Erlbaum.

Chater, N., Tenenbaum, J. B. & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, *10*(7), 287–291.

Church, J. A., Balota, D. A., Petersen, S. E. & Schlaggor, B. L. (2011). Manipulation of length and lexicality localizes the functional neuroanatomy of phonological processing in adult readers. *Journal of Cognitive Neuroscience*, *23*(6), 1475–1493.

Cohen, G. & Faulkner, D. (1986). Memory for proper names: Age differences in retrieval. *British Journal of Developmental Psychology*, *4*, 187–197.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral science (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Coltheart, M., Davelaar, E., Jonasson, J. T. & Besner, D. (1977). Access to the internal lexicon. In S. Dornick (Ed.), *Attention and performance* (Vol. VI, p. 535-556). Hillsdale, New Jersey: Erlbaum.

Coltheart, M. & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–258.

Coltheart, V., Laxon, V. J. & Keating, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology*, *79*, 1–12.

Cornelissen, P. L., Tarkiainen, A. & Salmelin, R. (2003). Cortical effects of shifting letter position in letter strings of varying length. *Journal of Cognitive Neuroscience*, *15*(5), 731–746.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Dabrowska, E. (2000). From formula to schema: the acquisition of english questions. *Cognitive Linguistics*, *11*(1/2), 83–102.

Daelemans, W. & Bosch, A. Van den. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, W., Bosch, A. Van den & Weijters, A. (1997). IGTree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, *11*, 407-423.

Daelemans, W., Zavrel, J., Sloot, K. Van der & Bosch, A. Van den. (2007). *TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.1* (Technical Report No. ILK 07-07). Computational Linguistics Tilburg University.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill Publishing Company.

Daw, N. & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593–620.

Dehaene, S., Cohen, L., Sigman, M. & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Science*, *9*, 335–341.

Dehaene, S., Le Clec, H. G., Poline, J. B., Le Bihan, D. & Cohen, L. (2002). The visual word form area: a prelexical representation of visual words in the fusiform gyrus. *Neuroreport*, *13*, 321–325.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M. & Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. *Brain and Language*, *81*, 555-567.

De Jong, N. H., Schreuder, R. & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*, 329-365.

Dell, G. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, *93*, 283-321.

Derouesne, J. & Beauvois, M. F. (1985). The phonemic stage in the nonlexical reading process: Evidence from a case of phonological alexia. In K. Patterson, J. C. Marschall & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 399–458). London: Erlbaum.

De Saussure, F. (1966). *Course in general linguistics*. New York: McGraw.

Diependaele, K., Sandra, D. & Grainger, J. (2005). Masked cross-modal morphological priming: Unravelling morpho-orthographic and morpho-semantic influences in early word recognition. *Language and Cognitive Processes*, *20*, 75–114.

Dijkstra, T., Prado Martín, F. Moscoso del, Schulpen, B., Schreuder, R. & Baayen, R. (2005). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes*, *20*, 7–41.

Doyle, A. C. (1982). *The Adventures of Sherlock Holmes*. England: George Newnes Ltd.

Duñabeitia, J. A., Perea, M. & Carreiras, M. (2007). The role of frequency constituents in compound words: Evidence from Basque and Spanish. *Psychonomic Bulletin & Review*, *14*, 1171–1176.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Engbert, R., Nuthmann, A., Richter, E. & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777–813.

Facoetti, A., Zorzi, M., Cestnick, L., Lorusso, M. L., Molteni, M. & Paganoni, P. (2006). The relationship between visuo-spatial attention and nonword reading in developmental dyslexia. *Cognitive Neuropsychology*, *23*, 841–855.

Fiez, J. A., Balota, D. A., Raichle, M. & Peterson, S. E. (1999). Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading. *Neuron*, *24*, 205–218.

Fodor, J. & Pylyshyn, F. (1988). Connectionism and cognitive architecture. *Cognition*, *28*(1–2), 3–71.

Forster, K. I. & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627-635.

Francis, W. N. & Kucera, H. (1979). Brown corpus manual. *Brown University Department of Linguistics*.

Frederiksen, J. & Kroll, J. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 361–379.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378.

Friedman, L. & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, *59*, 127–136.

Friston, K. (2005). A theory of cortical responses. *Phil. Trans. R. Soc. B*, *360*, 815–836.

Gagné, C. & Shoben, E. J. (1997). The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 71–87.

Gagné, C. L. & Spalding, T. L. (2014). Conceptual composition: The role of relational competition in the comprehension of modifier-noun phrases and noun-noun compounds. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 97–130). New York: Elsevier.

Gallistel, C. (2003). Conditioning from an information perspective. *Behavioural Processes*, *62*, 89–101.

Giraudo, H. & Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. *Psychonomic Bulletin and Review*, *8*, 127–131.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674–691.

Grabner, R. H., Brunner, C., Leeb, R., Neuper, C. & Pfurtscheller, G. (2007). Event-related EEG theta and alpha band oscillatory responses during language translation. *Brain research bulletin*, *72*(1), 57-65.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, *29*, 228–244.

Grindrod, C. M., Bilenko, N. Y., Myers, E. B. & Blumstein, S. E. (2008). The role of the left inferior frontal gyrus in implicit semantic competition and selection: An event-related fMRI study. *Brain Research*, *1229*, 167-178.

Hagoort, P., Indefrey, P., Brown, C., Herzog, H., Steinmetz, H. & Seitz, R. (1999). The neural circuitry involved in the reading of german words and pseudowords: A pet study. *Journal of Cognitive Neuroscience*, *11*, 383–398.

Han, J. & Kamber, M. (2000). Data mining: Concepts and techniques. *The Morgan Kaufman Series in Data Management Systems*.

Harm, M. W. & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491-528.

Harm, M. W. & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Hastie, T. & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, *1(3)*, 297-318.

Hauk, O., Davis, M., Ford, M., Pulvermüller, F. & Marslen-Wilson, W. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*, 1383–1400.

Henderson, L. (1982). *Orthography and word recognition in reading.* London: Academic Press.

Hendriks, W. & Kolk, H. (1997). Strategic control in developmental dyslexia. *Cognitive Neuropsychology*, *14*(3), 312–366.

Hendrix, P. (2008). *Electrophysiological effects in language production: a picture naming study using generalized additive modeling.* MSc thesis, Radboud University, Nijmegen, The Netherlands.

Herbster, A., Mintun, M. A., Nebes, R. & T., B. J. (1997). Regional cerebral blood flow during word and nonword reading. *Human Brain Mapping*, *5*, 84–92.

Hillyard, S. & Picton, T. (1987). Electrophysiology of cognition. *Handbook of physiology*, *5*, 519–584.

Hollerman, J. R. & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309.

Houghton, G. & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive Neuropsychology*, *20*, 115–162.

Hsu, A. S., Chater, N. & Vitányi, P. (2010). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Manuscript submitted for publication*.

Hyönä, J. & Pollatsek, A. (1998). Reading finnish compound words: Eye fixations are affected by component morphemes. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1612-1627.

Indefrey, P., Kleinschmidt, A., Merboldt, K. D., Kruger, G., Brown, C., Hagoort, P. et al. (1997). Equivalent responses to lexical and nonlexical visual stimuli in occipital cortex: A functional magnetic resonance imaging study. *Neuroimage*, *5*, 78–81.

Jacobs, A. M. & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334.

Janssen, N., Bi, Y. & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes*, *23*(7–8), 1191–1223.

Jared, D. (1997). Spelling-sound consistency affects the naming of high frequency words. *Journal of Memory and Language*, *36*, 505–529.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*, 723–750.

Jared, D., McRae, K. & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*, 687–715.

Jared, D. & Seidenberg, M. (1990). Naming multisyllabic words. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 92–105.

Jarema, G., Busson, C., Nikolova, R., Tsapkini, K. & Libben, G. (1999). Processing compounds: A cross-linguistic study. *Brain and Language*, *68*, 362–369.

Jobard, G., Crivello, F. & Tzourio-Mazoyer, N. (2003). Evaluation of the dual route theory of reading: a meta-analysis of 35 neuroimaging studies. *Neuroimage*, 693–712.

Juhasz, B. (2007). The influence of semantic transparency on eye movements during english compound word recognition. In R. Van Gompel, M. Fischer, W. Murray & R. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 373–389). Amsterdam: Elsevier.

Juhasz, B. & Berkowitz, R. (2011). Effects of morphological families on english compound word recognition: A multitask investigation. *Language and Cognitive Processes*, *26*(4), 653–682.

Juhasz, B., Starr, M., Inhoff, A. & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from lexical decision, naming, and eye fixations. *British Journal of Psychology*, *94*, 223–244.

Juphard, A., Vidal, J. R., Perrone-Bertolatti, M., Minotti, L., Kahone, P., Lachaux, J. P. et al. (2011). Direct evidence for two different neural mechanisms for reading familiar and unfamiliar words: An intra-cerebral EEG study. *Frontiers of Human Neuroscience*, *5*, 110.

Kan, I. P. & Thompson-Schill, S. L. (2004). Effect of name agreement on prefrontal activity during overt and covert picture naming. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 43-57.

Kennedy, A. (2003). *The Dundee Corpus.* [CD-ROM].

Klatt, D. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Kliegl, R., Nuthmann, A. & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*, 12-35.

Kryuchkova, T., Tucker, B. V., Wurm, L. & Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain and Language*, *122*(2), 81–91.

Kuperman, V., Bertram, R. & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*, 1089–1132.

Kuperman, V., Bertram, R. & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*, 83-97.

Kuperman, V., Schreuder, R., Bertram, R. & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, *35*, 876–895.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-38.

Libben, G. (2005). Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics*, *50*, 267–283.

Libben, G. (2006). Why study compound processing? In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 1–23). Oxford: Oxford University Press.

Libben, G., Gibson, M., Yoon, Y. & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, *84*, 50–64.

Marelli, M. & Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, *66*(4), 644–664.

Marmurek, H. H. C. & Kwantes, P. J. (1996). Reading words and wirds: Phonology and lexical access. *Quarterly Journal of Experimental Psychology*, *49*, 696–714.

McCann, R. & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 14–24.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. an account of the basic findings. *Psychological Review*, *88*, 375-407.

Milham, M. P., Banich, M. T. & Barad, V. (2003). Competition for priority in processing increases prefrontal cortex's involvement in top-down control: An event-related fMRI study of the stroop task. *Cognitive Brain Research*, *17*, 212-222.

Milin, P., Filipović Durdević, D. & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 50–64.

Milin, P., Kuperman, V., Kostić, A. & Baayen, R. (2009). Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition* (pp. 214–252). Oxford: Oxford University Press.

Milin, P., Ramscar, M., Coch, K., Feldman, L. & Baayen, R. H. (2015). Cornering segmentation: the perspective from discriminative learning. *Manuscript*.

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, *3*, 235-312.

Miller, R. R., Barnet, R. C. & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*(3), 363–386.

Monsell, S. (1985). Repetition and the lexicon. In S. Monsell (Ed.), *Progress in the psychology of language (Vol. 2)*. London: Erlbaum.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165-178.

Moscoso del Prado Martín, F. (2003). *Paradigmatic effects in morphological processing: Computational and cross-linguistic experimental studies*. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R. & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1271–1278.

Mulatti, C., Reynolds, M. G. & Besner, D. (2006). Neighborhood effects in reading aloud: New findings and new challenges for computational models. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 799–810.

Murre, J. M. J., Phaf, R. H. & Wolters, G. (1992). Calm: Categorizing and learning module. *Neural Networks*, *5*, 55–82.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189-234.

Norris, D. G. & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Novick, J., Trueswell, J. & Thompson-Schill, S. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, *4*(10), 906-924.

O'Regan, J. K. (1992). Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 333–354). Berlin, Germany: Springer-Verlag.

O'Regan, J. K. & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology*, *18*, 185–197.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Science*, *2*, 455–462.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural Computation*, 1199–1242.

Paap, K. R., McDonald, J. E., Schvaneveldt, R. W. & Noel, R. W. (1987). Frequency and pronounceability in visually presented naming and lexical decision tasks. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 221–243). Hillsdale NJ: Erlbaum.

Paap, K. R. & Noel, R. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research*, *53*, 13-24.

Patterson, K. & Behrmann, M. (1997). Frequency and consistency effects in a pure surface dyslexic patient. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 1217–1231.

Paulescu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N. & Cappa, S. (2000). A cultural effect on brain function. *Nature: Neuroscience*, *3*, 91–96.

Penolazzi, B., Hauk, O. & Pulvermüller, F. (2007). Early lexical access and semantic context integration as revealed by event-related brain potentials. *Biological Psychology*, *74*(3), 374–388.

Perry, C., Ziegler, J. & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review*, *114*(2), 273–315.

Perry, C., Ziegler, J. C. & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*(2), 106–151.

Pham, H. & Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience. In press*.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Pollatsek, A. & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes*, *20*, 261–290.

Pollatsek, A., Hyönä, J. & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human, Perception and Performance*, *26*, 820–833.

Price, C. J., Wise, R. J. S. & Frackowiak, R. S. J. (1996). Demonstrating the implicit processing of visually presented words and pseudowords. *Cerebral Cortex*, *6*, 62–70.

Ramscar, M., Dye, M., Gustafson, J. W. & Klein, J. (2013). Dual routes to cognitive flexibility: Learning and response-conflict resolution in the dimensional change card sort task. *Child Development*, *84*(4), 1308–1323.

Ramscar, M., Dye, M. & McCauley, S. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4), 760–793.

Ramscar, M., Hendrix, P., Love, B. & Baayen, H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, *8*(3), 450–481.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P. & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*, 5–42.

Ramscar, M., Smith, A. H., Dye, M., Futrell, R., Hendrix, P., Baayen, R. H. et al. (2013). The 'universal' structure of name grammars and the impact of social engineering on the evolution of natural information systems. *Proceedings of the 35th Meeting of the Cognitive Science Society*.

Ramscar, M. & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(7), 909–957.

Rastle, K. & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 482–503.

Rastle, K., Harrington, J. & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339–1362.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*(3), 618–660.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception*, *8*, 21–30.

Rayner, K., Well, A., Pollatsek, A. & Bertera, J. (1982). The availability of useful information to the right of fixation in reading. *Perception and Psychophysics*, *31*, 537–550.

Rescorla, R. & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Reynolds, M. & Besner, D. (2005). Basic processes in reading: A critical review of pseudohomophone effects in reading aloud and a new computational account. *Psychonomic Bulletin and Review*, *12*, 622–646.

Richardson, J. (1976). The effects of stimulus attributes on latency of word recognition. *British Journal of Psychology*, *67*, 315–325.

Ridgeway, G. (2015). gbm: Generalized boosted regression models [Computer software manual]. Available from `http://CRAN.R-project.org/package=gbm` (R package version 2.1.1)

Roberts, M. A., Rastle, K., Coltheart, M. & Besner, D. (2003). When parallel processing in visual word recognition is not enough: New evidence from naming. *Psychonomic Bulletin and Review*, *10*, 405–414.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Rosiers, G. Des & Ivison, D. (1986). Paired associate learning: Normative data for differences between high and low associate word pairs. *Journal of Clinical Experimental Neuropsychology*, *8*, 637–642.

Rossion, B., Schiltz, C. & Crommelinck, M. (2003). The functionally defined right occipital and fusiform "face areas" discriminate novel from visually familiar faces. *NeuroImage*, *19*, 877–883.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (p. 318-364). Cambridge, Mass.: The MIT Press.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing. explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (p. 216-271). Cambridge, Mass.: The MIT Press.

Rumelhart, D. E. & Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review*, *81*, 99-118.

Rumsey, J., Horwitz, B., Donohue, B., Nace, K., Maisog, J. & Andreason, P. J. (1997). Phonological and orthographic components of word recognition: A PET-rCBf study. *Brain*, *120*, 739–759.

Scarborough, D. L., Cortese, C. & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1-17.

Schreuder, R. & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (p. 131-154). Hillsdale, New Jersey: Lawrence Erlbaum.

Schreuder, R. & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Seidenberg, M. S. (2006). Connectionist models of reading. In G. Gaskell (Ed.), *The oxford handbook of psycholinguistics* (pp. 235–250). Oxford: University Press.

Seidenberg, M. S. & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*(9), 353-361.

Seidenberg, M. S. & McClelland, J. (1990). More words but still no lexicon: Reply to Besner et al. *Psychological Review*, *97*(3), 447-452.

Seidenberg, M. S. & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psycholgical Review*, *96*, 523–568.

Seidenberg, M. S., Petersen, A., MacDonald, M. C. & Plaut, D. C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 48–72.

Seidenberg, M. S. & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science*, *9*, 234–237.

Seidenberg, M. S. & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews. (Ed.), *From inkmarks to ideas: Current issues in lexical processing.* Hove, UK: Psychology Press.

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J. L. & McRae, K. (1994). Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1177–1196.

Sereno, S. C., Rayner, K. & Posner, M. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport*, *9(10)*, 2195-2200.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423.

Shaoul, C., Arppe, A., Hendrix, P., Milin, P. & Baayen, R. H. (2013). ndl: Naive discriminative learning [Computer software manual]. (R package version 0.2.14)

Shaoul, C., Westbury, C. F. & Baayen, R. H. (2013). The subjective frequency of word n-grams. *Psihologija*, *46*, 497–537.

Siegel, S. & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321.

Siyanova-Chanturia, A., Conklin, K. & Van Heuven, W. (2011). Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multi-word sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776–784.

Snyder, H. R., Feigenson, K. & Thompson-Schill, S. L. (2007). Prefrontal cortical response to conflict during semantic and phonological tasks. *Journal of Cognitive Neuroscience*, *19*, 761-775.

Spieler, D. H. & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *6*, 411–416.

Spreen, O. & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary.* Oxford, UK: Oxford University Press.

Strijkers, K., A., C. & G., T. (2010). Tracking lexical access in speech production: electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex*, *20*(4), 912–928.

Sun, C. C., Hendrix, P., Baayen, R. H. & Ramcar, M. (2015). The price of knowledge: bilingual paired associate learning. *Manuscript*.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition*, *7*, 263-272.

Taft, M. (1991). *Reading and the mental lexicon.* Hove, U.K.: Lawrence Erlbaum.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, *57A*, 745–765.

Taft, M. & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638-647.

Taft, M. & Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*, 607-620.

Taft, M. & Russell, B. (1992). Pseudohomophone naming and the word frequency effect. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *45*, 51–71.

Taraban, R. & McClelland, J. L. (1987). Consistency effects in word recognition. *Journal of Memory and Language*, *26*, 608–631.

Tarkiainen, A., Helenius, P., Hansen, P. C., Cornelissen, P. L. & Salmelin, R. (1999). Dynamics of letter string perception in the human occipotemporal cortex. *Brain*, *122*, 2119–2131.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences*, *94*, 14792-14797.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, Mass.: Harvard University Press.

Treiman, R., Mullennix, J., Bijeljac-Babic, R. & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of english orthography. *Journal of Experimental Psychology: General*, *124*, 107–136.

Tremblay, A. (2010). icaocularcorrection: Independent Components Analysis (ICA) based eye-movement correction [Computer software manual]. Available from `http://CRAN.R-project.org/package=icaOcularCorrection` (R package version 1.2)

Tremblay, A. & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

Tremblay, A., Baayen, R. H., Derwing, B., Libben, G., Tucker, B. & Westbury, C. (2011). Empirical evidence for an inflationist lexicon. *Proceedings of the Annual Meeting of the Linguistics Society of America*.

Van Jaarsveld, H. J. & Rattink, G. E. (1988). Frequency effects in the processing of lexicalized and novel nominal compounds. *Journal of Psycholinguistic Research*, *17*, 447-473.

Van Turennout, M., Bielamowicz, L. & Martin, A. (2003). Modulation of neural activity during object naming: Effects of time and practice. *Cerebral Cortex*, *13*, 381–391.

Vinckier, F., S, D., Jobert, A., Dubus, J. P., Sigman, M. & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*, 143–156.

Waelti, P., Dickinson, A. & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43–48.

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *50*, 439–456.

Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*, 1–15.

Widrow, G. & Hoff, M. E. (1960). Adaptive switching circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention record, part 4* (pp. 96–104). New York: Institute of Radio Engineers.

Wilson, T. W., Leuthold, A. C., Lewis, S. M., Georgopoulos, A. P. & Pardo, P. J. (2005). The time and space of lexicality: a neuromagnetic view. *Experimental Brain Research*, *162*(1), 1–13.

Wood, S. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wurm, L. H. & Fisicaro, S. A. (2014). What residualizing predictors in regression analysis does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.

Wydell, T., Vuorinen, T., Helenius, P. & Salmelin, R. (2003). Neural correlates of letter-string length and lexicality during reading in a regular orthography. *Journal of Cognitie Neuroscience*, *15*, 1052–1062.

Xu, B., Grafman, J., Gaillard, W. D., Ishii, K., Vega-Bermudez, F. & Pietrini, P. (2001). Conjoint and extended neural networks for the computation of speech codes: The neural basis of selective impairment in reading words and pseudowords. *Cerebral Cortex*, *11*, 267–277.

Yarkoni, T., Balota, D. A. & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Yeung, N., Cohen, J. D. & Botvinick, M. M. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*, 931–959.

Zevin, J. D. & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming. *Journal of Memory and Language*, *54*, 145–160.

Ziegler, J. C. & Perry, C. (1998). No more problems in Coltheart's neighborhood: Resolving neighborhood conflicts in the lexical decision task. *Cognition*, *68*, B53–B62.

Ziegler, J. C., Perry, C., Jacobs, A. M. & Braun, M. (2001). Identical words are read differently in different languages. *Psychological Science*, *12*, 379-384.

Zorzi, M. (1999). *The connectionist dual-process model: Development, skilled performance, and breakdowns of processing in oral reading.* Doctoral dissertation, University of Trieste, Trieste, Italy.

Zorzi, M., Houghton, G. & Butterworth, B. (1998a). The development of spelling-sound relationships in a model of phonological reading. *Language and Cognitive Processes*, *13*, 337–371.

Zorzi, M., Houghton, G. & Butterworth, B. (1998b). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1131–1161.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Language and Cognitive Processes*, *9*, 341-368.

Summary

Every day, we use language to communicate about the world around us in a seemingly effortless manner. Without any significant problems, we understand others and others understand us as we convey information about countless objects and events in this world. Rarely, if ever, do we ask ourselves the question "How is this possible?".

Psycholinguistics is a field of research that tries to answer this and many more questions about the human language processing system. Typically, data are collected through experiments in a lab, in which participants are asked to complete a linguistic task while their behavior is tracked in one form or another. Subsequently, the data from these experiments are analyzed in an attempt to gain new insights into linguistic processing.

Oftentimes, psycholinguistic researchers investigate the effects of lexical distributional variables on behavioral measures of language processing. Lexical distributional variables are measures that describe the distributional properties of a linguistic stimulus, such as the frequency of occurrence of a word or the number of words that are similar in form to a word (e.g., "life" is similar to "wife"). The effects of lexical distributional variables inform us about *which* properties of linguistic stimuli influence language processing. They provide no information, however, about *why* these properties - and not others - are important.

While lexical distributional variables provide different higher-level windows on the language processing system, this dissertation is an attempt to describe the language processing system itself. Unlike analyses and linguistic models that are based on lexical distributional variables, it takes into account the role of learning. The point of departure for the

analyses of linguistic data sets presented here is a simple general-purpose probabilistic learning algorithm (cf. Chater et al., 2006; Hsu et al., 2010, see also Baayen et al., 2011): the Rescorla-Wagner equations (Rescorla & Wagner, 1972). As a mathematical formalization of discrimination learning, the Rescorla-Wagner equations describe how people learn to respond differently to different stimuli, be they linguistic or non-linguistic in nature.

The Rescorla-Wagner equations describe a two-layer network model, in which both input units and outcomes are symbols. In the work presented here, these symbols are linguistic units, such as letters, phonemes or words. As such, the symbolic approach used here stands in contrast to sub-symbolic approaches, in which linguistic units are represented as activation patterns over non-symbolic units (which, at a lower level of granularity, are again symbolic). Symbolic models are an oversimplification of a more complex neurobiological reality (as are many implementations of sub-symbolic models), but provide highly competitive performance and an increased interpretability as compared to sub-symbolic models of language processing.

More precisely, the foundation of the work presented here was laid down in Baayen et al. (2011), who describe an implementation of – the equilibrium equations for (Danks, 2003) – the Rescorla-Wagner equations in a model for silent reading. Given that the associations between input units and outcomes were estimated independently for each outcome - an assumption similar to the independence assumption in a statistical classification technique referred to as Naive Bayesian Classifiers - Baayen et al. (2011) refer to their model as the Naive Discriminative Reader (NDR). The NDR model accounted for a wide range of effects documented in the experimental reading literature.

The NDR model, however, was limited to silent reading. To truly gauge the potential of a computational approach to language processing, it is pivotal to investigate its performance across a variety of experimental tasks and the behavioral measures of language processing obtained through these tasks. This dissertation provides a more extensive evaluation of the possibilities offered by a discrimination learning ap-

proach to language processing, by looking at the explanatory power of discrimination learning networks in three different experimental tasks and for three different dependent variables.

First, this thesis presents an extension of the NDR model for silent reading to reading aloud. The resulting Naive Discriminative Reading Aloud (NDR_a) model consists of two discrimination learning networks. The first network maps orthographic features onto lexico-semantic representations, similar to the discrimination learning network for silent reading described in Baayen et al. (2011). The second network maps lexico-semantic representations onto phonological features.

Existing models of reading aloud typically consist of two routes: a lexical route in which the orthography-to-phonology mapping is mediated by lexico-semantic representations and a sub-lexical route, in which orthographic units are mapped directly onto phonological units. By contrast, a single lexical architecture is responsible for both word and non-word naming in the NDR_a model presented here.

In word reading, the orthographic presentation of the target word (e.g., "life") activates the lexico-semantic representation of the target word, as well as the lexico-semantic representations of orthographically similar words (e.g., "wife", "knife"). These lexico-semantic representations then activate phonological units, which allow for the pronunciations of the target word. For a non-word (e.g., "kife"), no lexico-semantic representation exists. The activation of phonological units, therefore, is driven exclusively by the activation of orthographic neighbors of the non-word (e.g., "life", "wife", "knife").

An extensive evaluation of the NDR_a model demonstrates that the single-route architecture of the NDR_a model is capable of capturing a wide range of effects documented in the reading aloud literature, both for words and for non-words, including linear and non-linear effects of neighborhood density measures and the consistency of the orthography to phonology mapping, as well as a hitherto unobserved effects of non-word frequency. Despite its much more parsimonious model architecture, the overall performance of the NDR_a model is highly similar to that of a state-of-the-art dual-route model of reading aloud (see Perry et al.,

2007). Furthermore, the addition of a sub-lexical route architecture does not further improve the performance of the NDR_a model. When using a discrimination learning approach, therefore, a single lexical route is sufficient to provide a highly competitive account of language processing in the reading aloud task.

The second test-case for a discrimination learning approach presented here is an investigation of the eye fixation patterns during noun-noun compound reading in a new large-scale corpus of eye movements during natural discourse reading, the Edmonton-Tübingen eye-tracking corpus (ET corpus). An analysis using distributional lexical variables indicates that the fixation patterns on compounds in the ET corpus do not fit straightforwardly with existing sub-lexical (constituent access precedes full-form access), supra-lexical (full-form access precedes constituent access) or dual-route (a holistic and a decompositional route are pursued in parallel) models of compound reading. An analysis using predictors derived from two naive discrimination learning (henceforth NDL) networks sheds further light on the processes that drive compound reading and suggests that compound reading is perhaps better thought of as an attempt to activate the lexico-semantic information associated with a compound given all information available to the reader (cf. maximization of opportunities, Libben, 2006).

Over 60% of the time, a single fixation on a compound suffices. During single fixations, readers fixate far enough into a compound to make all orthographic features of the compound available. These orthographic features then activate all lexico-semantic information associated with a compound: first-and-only fixation durations are co-determined by an integrative measure of the bottom-up support for the lexico-semantic representations of not only the compound as a whole, but also the modifier and the head.

Nearly 40% of the time, readers need a second fixation to successfully process a compound. An important cause of additional fixations is a suboptimal fixation position during the first fixation. As a result, not all orthographic features of a compound are available to the reader during first-of-many fixations. Accordingly, the NDL measure that proved most

predictive for first-and-only fixation duration is the activation of the lexico-semantic representation of the modifier given the first orthographic trigram of a compound only.

In the analysis using lexical distributional variables, both first-and-only and second fixation durations were influenced by the frequency of the compound. On the basis of a lexical predictor analysis one might therefore be tempted to conclude that the processes underlying single fixations and second fixations are similar. The NDL analysis of the compound reading data, however, demonstrates that the frequency effects for first-and-only fixations and second fixations are qualitatively different. Whereas first-and-only fixation durations are characterized by the bottom-up support for the lexico-semantic information associated with a compound, second fixation durations are influenced by the out-of-context, a priori probability of a compound. In other words: during second fixations readers resort to a top-down "best guess" strategy.

Overall, the explanatory power of lexical distributional variables and that of NDL measures was highly similar. As for the reaction times in the reading aloud task, therefore, the NDL framework offers a highly competitive perspective on eye fixation patterns during compound reading. As demonstrated above, however, the NDL measures provide more detailed and more differentiated insights into the processes that underlie these fixation patterns.

Third, the perhaps most stringent test of the potential of discrimination learning in this dissertation is an exploration of the explanatory power of NDL measures for the electroencephalographic correlates of language processing in a primed picture naming task, as gauged through the ERP signal following picture onset. In this primed picture naming task, participants were presented with preposition plus definite article primes (e.g., "on the") and target pictures of concrete nouns (e.g., "STRAWBERRY"). The data for this primed picture naming experiment show an effect for preposition frequency, theta range oscillations for word frequency and constructional prototypicality, and a prolonged near-linear effect for phrase frequency.

An NDL analysis of the ERP data shows an effect of the bottom-up support for preposition that qualitatively and topographically resembles the effect of preposition frequency. Furthermore, the effect for the bottom-up support for the target word shows some similarities to the effect of word frequency. Most similar to the effect of word frequency, however, is the effect of top-down information about the a priori probability of the target word. Again, therefore, the NDL analysis provides more detailed insight into the nature of a frequency effect.

The quantitive performance of the NDL measures in explaining the ERP signal after picture onset is at least as good as the performance of lexical distributional variables. In a regression analysis NDL measures and lexical predictors explain a very similar amount of the variance in the ERP signal. According to an analysis with a tree-based machine learning technique, however, NDL measures significantly outperform lexical distributional variables.

In conclusion, this dissertation assesses the potential of discrimination learning as a perspective on the adult language processing system. Across three experimental tasks and three behavioral measures of language processing, discrimination learning shows highly competitive performance as compared to existing analysis techniques and models of language processing, and allows for new or more refined insights into the systemic properties that drive language processing through a general-purpose learning mechanism that is remarkably simple and transparent.

Zusammenfassung

Täglich verwenden wir Sprache scheinbar mühelos um uns über unsere Umwelt auszutauschen. Ohne größere Probleme verstehen wir Andere und Andere verstehen uns, wenn wir Informationen über zahllose Objekte und Ereignisse in der Welt übermitteln. Selten fragen wir uns: “Wie ist das möglich?”.

Psycholinguistik ist ein Forschungsgebiet, welches diese und viele weitere Fragen über das menschliche Sprachverarbeitungssystem zu beantworten versucht. Typischerweise geschieht das Sammeln von Daten durch Experimente in einem Labor, wobei Versuchspersonen eine linguistische Aufgabe durchführen, während auf die eine oder andere Weise ihr Verhalten beobachtet wird. Anschließend werden die Daten aus diesen Experimenten ausgewertet, mit dem Ziel neue Einblicke in die linguistische Verarbeitung zu gewinnen.

Oft untersuchen Psycholinguisten die Auswirkungen lexikalischer Variablen auf Verhaltensmaße der Sprachverarbeitung. Lexikalische Variablen sind Messgrößen, welche die Verteilungseigenschaften eines linguistischen Stimulus beschreiben, so wie die Worthäufigkeit oder die Anzahl an Wörtern, die von der Form her ähnlich sind (bspw. “life” und “wife”). Sie werden deshalb auch lexikalische Verteilungsvariablen genannt. Die Auswirkungen lexikalischer Verteilungsvariablen verraten uns, *welche* Eigenschaften linguistischer Stimuli einen Einfluss auf die Sprachverarbeitung haben. Allerdings geben sie keinerlei Auskunft darüber, *weshalb* diese Eigenschaften – und nicht andere – wichtig sind.

Während lexikalische Verteilungsvariablen verschiedene Einblicke in die übergeordneten Prozesse der Sprachverarbeitung ermöglichen, versucht diese Dissertation das Sprachverarbeitungssystem an sich zu mod-

ellieren. Im Gegensatz zu Analysen und linguistischen Modellen, die auf lexikalischen Verteilungsvariablen beruhen, berücksichtigt diese Arbeit die Rolle des Lernens. Der Ausgangspunkt für die Analyse linguistischer Datensätze, wie sie hier vorgestellt wird, ist ein einfacher universeller probabilistischer Lernalgorithmus (vgl. Chater et al., 2006; Hsu et al., 2010, siehe auch Baayen et al., 2011): die Rescorla-Wagner-Gleichungen (Rescorla & Wagner, 1972). Als mathematische Formalisierung des "Discrimination Learning" beschreiben die Rescorla-Wagner-Gleichungen, wie Menschen lernen, unterschiedlich auf unterschiedliche Stimuli zu reagieren.

Die Rescorla-Wagner-Gleichungen beschreiben ein zweischichtiges Netzwerk-Modell, in welchem sowohl die Eingabe als auch die Ausgabe aus diskreten Symbolen besteht. In der hier vorgestellten Arbeit sind diese Symbole linguistische Einheiten wie Buchstaben, Phoneme oder Wörter. Damit steht der hier verwendete symbolische Ansatz im Kontrast zu subsymbolischen Ansätzen, in welchen linguistische Einheiten durch Aktivitätsmuster über nicht-symbolischen Einheiten repräsentiert sind (diese nicht-symbolischen Einheiten können auf niedrigerer Ebene wieder als symbolisch angesehen werden). Symbolische Modelle sind eine starke Vereinfachung einer komplexeren neurobiologischen Realität (wie viele Implementierungen von subsymbolischen Modellen), bieten dafür aber eine gute Performance und eine direktere Interpretierbarkeit im Vergleich zu subsymbolischen Modellen.

Der Grundstein der hier vorgestellten Arbeit wurde durch Baayen et al. (2011) gelegt, die eine Implementierung der Gleichgewichtsgleichungen für die Rescorla-Wagner-Gleichungen (Danks, 2003) in einem Modell für stilles Lesen beschreiben. Baayen et al. (2011) nennen ihr Modell den "Naive Discriminative Reader" (NDR), in welchem vorausgesetzt wird, dass die Assoziationen zwischen Eingabe und Ausgabe unabhängig voneinander für jede Ausgabe geschätzt werden – eine Annahme ähnlich der Unabhängigkeitsannahme in einer statistischen Klassifizierungstechnik bekannt als Bayes-Klassifikator. Das NDR Modell erklärte eine große Auswahl an Ergebnissen aus der Literatur des experimentellen Lesens.

Allerdings war das NDR Modell auf stilles Lesen beschränkt. Um das Potential eines computergestützten Ansatzes für die Sprachverarbeitung wirklich einschätzen zu können, ist es entscheidend, dessen Leistung über eine Vielfalt von experimentellen Aufgaben und die dabei erlangten Verhaltensmaße der Sprachverarbeitung, zu betrachten. Diese Dissertation liefert eine umfassendere Einschätzung der Möglichkeiten, die ein "Discrimination Learning" Ansatz der Sprachverarbeitung bietet, indem sie die Aussagekraft von "Discrimination Learning" Netzwerken in drei unterschiedlichen Experimenten und für drei verschiedene abhängige Variablen betrachtet.

Zuerst stellt diese Arbeit eine Erweiterung des NDR Modells für stilles Lesen auf lautes Lesen vor. Das resultierende "Naive Discriminative Reading Aloud" (NDR_a) Modell besteht aus zwei "Discrimination Learning" Netzwerken. Das erste Netzwerk bildet orthographische Merkmale auf lexiko-semantische Repräsentationen ab, ähnlich dem "Discrimination Learning" Netzwerk für stilles Lesen des NDR Modells. Das zweite Netzwerk bildet lexiko-semantische Repräsentationen auf phonologische Merkmale ab.

Existierende Modelle des lauten Lesens bestehen typischerweise aus zwei Pfaden: einem lexikalischen Pfad, auf welchem die Abbildung von Orthographie auf Phonologie über lexiko-semantische Repräsentationen durchgeführt wird, und einem sublexikalischen Pfad, auf welchem orthographische Einheiten direkt auf phonologische Einheiten abgebildet werden. Im Gegensatz dazu verwendet das hier vorgestellte NDR_a Modell eine einzige lexikalische Architektur, die sowohl für Wort- als auch Nonwort-Benennung verantwortlich ist.

Beim Lesen von Wörtern aktiviert die orthographische Darstellung des Zielwortes (z. B. "life") sowohl die lexiko-semantische Repräsentation des Zielwortes als auch die lexiko-semantischen Repräsentationen von orthographisch ähnlichen Wörtern (z. B. "wife", "knife"). Diese lexiko-semantischen Repräsentationen aktivieren wiederum phonologische Einheiten, welche die Aussprache des Zielwortes ermöglichen. Für ein Nonwort (z. B. "kife") existiert keine lexiko-semantische Repräsentation. Die Aktivierung von phonologischen Einheiten wird daher ausschließlich

durch die Aktivierung von orthographischen Nachbarn des Nonwortes (z. B. “life”, “wife”, “knife”) angetrieben.

Eine umfassende Betrachtung des NDR$_a$ Modells zeigt, dass die Ein-Pfad-Architektur des NDR$_a$ Modells in der Lage ist eine Vielfalt von Ergebnissen aus der Literatur zu lautem Lesen zu erfassen. Dies gilt sowohl für Wörter als auch für Nonwörter und bezieht sich zum Beispiel auf Effekte von “neighborhood density” Maßen und Effekte des Zusammenspiels von Ortographie und Phonologie. Zusätzlich sagt das NDR$_a$ Modell bisher nicht berichtete Einflüsse von Nonwort-Häufigkeit vorher, welche auch in am Menschen gemessenen Daten gefunden werden. Trotz seiner sparsamen Modellarchitektur ist die quantitative Leistung des NDR$_a$ Modells der eines aktuellen Zwei-Pfad-Modells des lauten Lesens (see Perry et al., 2007) sehr ähnlich. Des Weiteren verbessert das Hinzufügen einer sublexikalischen Pfad-Architektur die Leistung des NDR$_a$ Modells nicht weiter. Daher ist, beim Verwenden eines “Discrimination Learning” Ansatzes, ein einziger lexikalischer Pfad ausreichend, um eine mit anderen aktuellen Ansätzen vergleichbare Beschreibung der Sprachverarbeitung in der Aufgabe des lauten Lesens zu bieten.

Eine zweite hier vorgestellte Anwendung des “Discrimination Learning” Ansatzes in der Linguistik ist eine Untersuchung der Augenfixationsmuster während des Lesens von Nomen-Nomen-Komposita, die innerhalb von längeren Prosatexten vorkamen. Die Eyetracking-Daten wurden während der hier vorgestellten Arbeit erhoben und bilden einen Teil des Edmonton-Tübingen eye-tracking corpus (ET corpus). Eine Analyse mit lexikalischen Verteilungsvariablen zeigt, dass die Fixationsmuster für Komposita im ET corpus nicht direkt mit bestehenden sublexikalischen (Teilwortzugriff vor Ganzwortzugriff), supralexikalischen (Ganzwortzugriff vor Teilwortzugriff) oder Zwei-Pfad-Modellen (ein ganzheitlicher und ein dekompositioneller Pfad werden gleichzeitig verfolgt) des Lesens von Komposita übereinstimmen. Eine Analyse mit Prädiktoren, die aus zwei “Naive Discrimination Learning” (NDL) Netzwerken abgeleitet werden, gibt weiteren Aufschluss über die Prozesse, welche das Lesen von Komposita lenken, und legt nahe, dass beim Lesen von Komposita alle verfügbaren Informationen vom Leser genutzt werden um die mit dem

Komposita assoziierte lexiko-semantischen Informationen zu aktivieren (vgl. maximization of opportunities, Libben, 2006).

Mehr als 60% der Zeit reicht eine einzige Fixation auf das Kompositum aus. Während dieser einzelnen Fixation fixieren Leser weit genug in ein Kompositum hinein um alle seine orthographischen Eigenschaften verfügbar zu machen. Diese orthographischen Eigenschaften wiederum aktivieren die gesamte lexiko-semantische Information, die mit einem Kompositum assoziiert ist: Dauern von "first-and-only" Fixationen sind mitbestimmt durch eine einheitliche Messung der Bottom-up Unterstützung der lexiko-semantischen Repräsentation sowohl des Kompositums als Ganzem, als auch seiner Teile.

In beinahe 40% der Fälle benötigen die Leser eine zweite Fixation um das Kompositum erfolgreich zu verarbeiten. Ein wichtiger Auslöser für zusätzliche Fixationen ist eine suboptimale Fixationsposition während der ersten Fixation. Infolgedessen stehen dem Leser bei "first-of-many" Fixationen nicht alle orthographischen Eigenschaften zur Verfügung. Dementsprechend hat bei "first-and-only" Fixationen die NDL Aktivierung der lexiko-semantischen Repräsentation des linken Teils die größte Vorhersagekraft, wenn ausschließlich das erste Trigram des Kompositums als Eingabe verwendet wird.

In der Analyse mit lexikalischen Verteilungsvariablen zeigt sich ein Einfluss der Worthäufigkeit des Kompositums sowohl auf die "first-and-only" Fixation als auch auf die zweite Fixation. In Anlehnung an die Analyse mit lexikalischen Verteilungsvariablen könnte man daher annehmen, dass die grundlegenden Prozesse von "first-and-only" Fixationen und zweiten Fixationen ähnlich seien. Die NDL Analyse der Komposita-Daten zeigt allerdings, dass sich die Worthäufigkeitseffekte für "first-and-only" Fixationen und zweite Fixationen qualitativ unterscheiden. Während "first-and-only" Fixationsdauern durch die Bottom-Up Unterstützung der mit einem Kompositum assoziierten, lexiko-semantischen Informationen charakterisiert sind, werden zweite Fixationen durch die unabhängige a priori Wahrscheinlichkeit eines Kompositums beeinflusst. Anders gesagt: Während der zweiten Fixation greifen die Leser auf eine Top-Down "best guess" Strategie zurück.

Allgemein ist die Vorhersagekraft von lexikalischen Verteilungsvariablen und die von NDL Werten ähnlich. Wie bei den Reaktionszeiten in der Aufgabe "lautes Lesen" schneidet die NDL Struktur auch bei der Vorhersage von Augenfixationsmustern beim Lesen von Komposita im Vergleich mit anderen aktuellen Ansätzen gut ab. Wie oben dargestellt, liefern die NDL Werte weit detailliertere und differenziertere Einblicke in die Prozesse, die diesen Fixationsmustern zugrunde liegen.

Der dritte, vielleicht strengste, Test des Potentials von "Discrimination Learning" in dieser Dissertation ist eine Untersuchung der Vorhersagekraft der NDL Werte für die elektroenzephalographischen Korrelate der Sprachverarbeitung in einem Priming-Experiment zur Bildbenennung. Als zu untersuchendes Korrelat wurde das, auf den Onset des Bildes folgende, ERP Signal ausgewählt und gemessen. In diesem Priming-Experiment wurde den Versuchspersonen als Prime eine Präposition zusammen mit einem bestimmten Artikel gezeigt (z. B. "on the"), gefolgt von einem Bild eines freigestellten Objektes (z. B. einer Erdbeere). Die Daten aus diesem Priming-Experiment zeigen einen Effekt der Präpositionshäufigkeit, einen anhaltenden, beinahe-linearen Effekt für die Phrasenhäufigkeit und oszillatorische Effekte der Worthäufigkeit und der "Constructional Prototypicality", welche beide die ERP-Aktivität im Theta-Bereich modulieren.

Eine NDL Analyse der ERP Daten zeigt einen Effekt der Bottom-Up Aktivierung der Präposition, der qualitativ und topographisch dem Effekt der Präpositionshäufigkeit ähnelt. Des Weiteren weist der Effekt der Bottom-Up Aktivierung des Zielworts einige Ähnlichkeiten zum Worthäufigkeitseffekt auf. Die stärkste Ähnlichkeit zum Worthäufigkeitseffekt findet sich allerdings beim Effekt der Top-down Information, genauer in der a priori Wahrscheinlichkeit des Zielwortes. Wiederum liefert die NDL Analyse deshalb einen detaillierteren Einblick in das Wesen eines Häufigkeitseffektes.

Die quantitative Leistung der NDL Prädiktoren beim Erklären des ERP Signals nach dem Erscheinen des Bildes ist mindestens so gut wie die Leistung lexikalischer Verteilungsvariablen. In einer Regressionsanalyse erklären NDL Prädiktoren und lexikalische Prädiktoren jeweils einen ähnlich großen Anteil der Varianz im ERP Signal. In einer Analyse, der

eine baumbasierte “Machine Learning” Technik zugrunde liegt, zeigen die NDL Prädiktoren aber eine signifikant bessere Vorhersagekraft als die lexikalischen Verteilungsvariablen.

Diese Dissertation zeigt das Potential des “Discrimination Learning” als Beschreibung des Sprachverarbeitungssystems von Erwachsenen auf. Drei Experimente und drei Verhaltensmessungen der Sprachverarbeitung zeigen, dass “Discrimination Learning” im Vergleich zu anderen Analysetechniken und Sprachverarbeitungsmodellen vergleichbare oder bessere Ergebnisse liefert. Darüber hinaus ermöglicht “Discrimination Learning” neue und detailliertere Einblicke in die zugrunde liegenden Eigenschaften des Sprachverarbeitungssystems. Dabei bedient sich der hier vorgestellte Ansatz eines bemerkenswert einfachen, transparenten und universellen Lernmechanismus.