Processing Lexical Bundles

Lensink, S.E.

Processing Lexical Bundles

Published by

LOT
Kloveniersburgwal 48
1012 CX Amsterdam
The Netherlands

phone: +31 20 525 2461
e-mail: lot@uva.nl
http://www.lotschool.nl

The illustration on the cover shows a fragment of an alluvial diagram of all trigrams from the text of this dissertation beginning with 'hearing', 'reading', and 'speaking', the three big themes of this dissertation. The full illustration is printed on the back side of the propositions. To create the image, the author used the *alluvial* package in R, and the code is available upon request.

ISBN: 978-94-6093-350-9
NUR: 616

Processing Lexical Bundles

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 4 juni 2020
klokke 15:00 uur

door

Saskia E. Lensink

geboren 7 mei 1989
te Woudenberg, Nederland

Promotores:	prof. dr. A. Verhagen prof. dr. N.O. Schiller
Promotiecommissie:	prof. dr. A. Backus (Tilburg University) dr. Y. Chen prof. dr. M.P.G.M. Mous prof. dr. P. Zwitserlood (Westfälische Wilhelms-Universität Münster)

To all the parts that make me whole.

Contents

Acknowledgements xi

1 Introduction **1**
1.1 The What 2
1.2 The Why 2
1.3 The How 4
1.3.1 Reading 4
1.3.2 Listening 7
1.3.3 Speaking 7
1.4 Quantifying processing 9
1.4.1 Statistical modeling 9
1.4.2 Computational modeling 13
1.5 This dissertation 15

2 Reading **17**
2.1 Introduction 18
2.1.1 Language processing in younger and older adults 19
2.1.2 Reading lexical bundles in younger and older adults . . 20
2.1.3 The present study 20
2.2 Materials and Methods 21
2.2.1 Participants 21
2.2.2 Stimuli 21
2.2.3 Procedure 22
2.2.4 Generalized Additive Mixed-Effects models 22
2.3 Results 23
2.3.1 First fixation durations 23
2.3.2 Second fixation durations 26
2.3.3 Number of fixations 29
2.4 General discussion 31
2.4.1 The Inverted Frequency Effect 32

2.4.2 Limitations and future directions 33
2.5 Conclusion 35

3 Listening **37**
3.1 Introduction 38
3.1.1 Previous work on the time course of multi-word unit processing 39
3.1.2 Current study 41
3.2 Materials and Methods 41
3.2.1 Participants 41
3.2.2 Stimuli 42
3.2.3 Procedure 43
3.2.4 EEG recordings 43
3.2.5 Conditional inference random forest analysis 44
3.3 ERP results 45
3.4 CForest modeling results 47
3.4.1 Results Conditional Inference Tree 48
3.4.2 Stage 1: Prediction and bottom-up information, 0-270 ms 49
3.4.3 Stage 2: Inhibition and competition, 270-402 ms 56
3.4.4 Stage 3: Lexical integration, 402-800 ms 58
3.5 General discussion 63

4 Reading and speaking **67**
4.1 Introduction 68
4.1.1 Including multi-word units in models of lexical access . 69
4.1.2 Computational modeling of multi-word units 70
4.2 NDL model 71
4.2.1 How the model works 71
4.2.2 NDL measures of lexical access 73
4.3 Generalized additive mixed models 76
4.4 Eye-tracking experiment 77
4.4.1 Materials 78
4.4.2 Design 78
4.4.3 Participants 79
4.4.4 Procedure 79
4.4.5 Analyses 79
4.4.6 First Fixation Durations 80
4.4.7 First Pass Reading Times 81
4.4.8 Number of fixations 84
4.4.9 Discussion eye-tracking data 86
4.5 Production experiment 88
4.5.1 Materials 88
4.5.2 Design 88
4.5.3 Participants 89
4.5.4 Procedure 89

4.5.5 Analyses . . . 89
4.5.6 Production onset latencies . . . 89
4.5.7 Production durations . . . 90
4.5.8 Discussion production data . . . 92
4.6 General discussion . . . 95
4.6.1 Lexical access of multi-word units . . . 96

5 Conclusion **99**
5.1 Reading . . . 100
5.2 Listening . . . 101
5.3 Reading and speaking . . . 102
5.4 Overall conclusions . . . 104
5.5 Useful models - an outlook . . . 105

Appendix A . . . 107

Appendix B . . . 113

Bibliography . . . 115

Nederlandse samenvatting . . . 129

About the author . . . 137

Acknowledgements

Completing a PhD thesis is often considered a solitary and lonely task. However, looking back at the past years, I realize that I was by no means lonely, and I am grateful for all the support and companionship I received from so many people.

First and foremost, I would like to thank my supervisors Arie Verhagen and Niels Schiller, without whose supervision and support this dissertation would not exist. Thank you for your creative insights, interesting discussions, and giving me the freedom to chose my own directions.

I am grateful to the Netherlands National Graduate School of Linguistics (LOT) for awarding me a scholarship to fund my PhD. Together with 'LOT-genoten' Jeroen Breteler, Sophie Villerius, Brechje Van Osch, LOT directors Frank Wijnen and Henriette de Swart, I had regular meetings where we discussed our progress, academia, ambitions, and life.

Martijn Wieling and Jacolien van Rij introduced me to statistical modeling. I am greatly indebted to them for this as they convinced me that I would be able to learn how to apply these models myself, even though I understood almost nothing at first. This has set me on a path of further developing my machine learning skills. I furthermore owe special thanks to Antoine Tremblay, who patiently introduced me to ensemble machine learning models, and whose enthusiasm and creative insights have been of invaluable importance in the research reported in Chapter 3.

My dissertation would have never been complete without the inspiring collaboration with the quantitative linguistics group at the university of Tübingen. I am especially grateful to Harald Baayen and Tineke Baayen Oudshoorn, who have invited me over to Tübingen a couple of times and made me feel right at home. It was an honor to have had the opportunity to work with Harald and applying some of my ideas within the NDL framework. Besides Harald and Tineke, I had a great time with Petar Milin, Tino Sering, Fabian Tomaschek, June Hendrix-Sun, Peter Hendrix, Karlina Denistia, and Denis Arnold.

In 2015, I spent the summer at the university of Chicago, where I attented

the LSA Summer Institute — one of the highlights of my time as a PhD candidate. I learned so much from the courses and teachers, and made great friends. Thank you Amos Teo, Grant Berry, Darcy Rose, Kailen Shantz, Bodo Winter, and Andy Wedel for the memories, inspiring talks, and many beers we shared. My time at Leiden was filled with laughs, ideas, and support of my colleagues from LUCL, in particular my roommates Maaike van Naerssen and Evelyn Bosma. Thank you for the great times, Jessy Nixon, Alex Reuneker, Véronique Verhagen, Elly Dutton, Bobby Ruijgrok, Olga Kepinska, Cesko Voeten, Aliza Glasbergen-Plas, Leticia Pablos Robles, Xander Vertegaal, Stefan Norbruis, Bob Schoemaker, Andreas Krogull, Martine Bruil, Kate Bellamy, María Carme Parafita Couto, Katja Lubina, and Marijn van 't Veer. Besides my colleagues, I am indebted to all people who participated in my experiments, all the students I had the pleasure to meet teaching different courses, Maxime Tulling, Lilian Ye, for their endless efforts in helping me collecting data, their critical questions, and creative energy, and especially Jos Pacilly without whom none of my experiments would have been a success.

Special thanks go to Rinus Verdonschot, with whom I had the pleasure to collaborate on several projects, even before I started my PhD. He introduced me to Katsuo Tamaoka, to whom I am grateful for the interesting conversations we had and the collaboration we set up. Thanks to Rinus I got the chance to share my knowledge on statistical modeling at Waseda University, which I consider another highlight of my PhD.

Finishing a PhD without the help and support of friends, and the liters of coffee, tea, wine, and beer we shared, is impossible. I am thankful to Marte Boonen, Pauline van Deursen, Majelle Verbraak, Anoushka Kloosterman, Laura Berveling, Mariska Veenendaal, Gert Poot, Daša Abtová, Stefan Penders, Joey Weidema, Fokelien Kootstra, Carlos Merchán Pulgarín, Pip Schuijt, Nina Ouddeken, Jurgen van den Heuvel, Benjamin Suchard, Hilde Gunnink, Bas Clercx, and Redmer Kronemeijer. My buddies both above and below water, especially Bart Braun, Nathalie Raats, Mariska Beijer, and Marit van Elsas, have helped me to relax and enjoy (underwater) life.

Wilma en Joop, dank jullie wel dat jullie altijd voor me klaarstonden, vooral op momenten dat het echt nodig was. Van jullie heb ik geleerd om hard te werken en nooit op te geven. Ik dacht dat een PhD afronden zwaar was, tot het noodlot toesloeg en we leerden wat écht moeilijk is. Papa, ik mis je elke dag.

Mi querido chanchi wawi, 'amortje' Diego Vásquez Villaseca, mil gracias por siempre estar conmigo, creyendo en mi, apoyandome en las buenas y en las malas. Tu amor y apoyo incondicional fueron un gran soporte para llegar a este logro. Te amo.

CHAPTER 1

Introduction

... the trick to being a scientist is to be open to using a wide variety of tools. — Leo Breiman (2001)

This is a dissertation about the processing of lexical bundles. What are lexical bundles, and why is it worth studying them? Why specifically study their processing, and how does one go about doing that? To answer these questions, I will discuss the what, why, and how of lexical bundle processing.

In this introductory chapter, I will discuss what a lexical bundle is, and why studying the way we process them provides linguists and psychologists with important insights into language and cognition in general. I will then move on to discuss the diverse experiments researchers have carried out to study the processing of lexical bundles and other types of multi-word units, the results they have found, and the conclusions they have drawn from their data.

This thesis focuses on the processing of lexical bundles, and does so by considering them from different angles: How do we read lexical bundles? Are there differences in processing between age groups? How do we process spoken lexical bundles? And how do we produce them? In answering these questions, I have employed both statistical and computational modeling, techniques which I will briefly introduce at the end of this chapter.

1.1 The What

Auctioneers and sportscasters are known for their ability to speak incredibly quickly. They speak fast and fluently in situations where they have to perform other tasks besides talking, such as keeping track of bids or balls. The way they achieve this extraordinary feat is by using a restricted set of common phrases and sentences over and over again (Kuiper, 1996). However, it is not only auctioneers and sportscasters who abundantly employ ready-made chunks of language — we all do.

Estimates differ, but in general it is assumed that about half of our spoken and written language consists of stock phrases, formulaic sequences, and common combinations of words (Biber et al., 1999; Erman and Warren, 2000). Because of their prevalence, a lot of researchers have investigated different types of multi-word units, each of them using different terms: Chunks, collocations, formulae, formulaic sequences, idioms, lexical bundles, lexical patterns, multi-word units, multi-word expressions, n-grams, prefabs, or superlemmas.

Multi-word units differ from each other on several dimensions: There are multi-word units that are non-compositional, such that one cannot derive their meaning from the meaning of their constituent words; there are multi-word units that are fully compositional. There are multi-word units that are frequent; there are multi-word units that are infrequent. Some multi-word units are very salient; others are not. After having surveyed experimental evidence on how different multi-word units are processed, Wray (2012) proposed a multi-dimensional space along which subtypes of multi-word units are distributed. See Figure 1.1 for a graphical representation.

As can be seen in Figure 1.1, idioms are typical instances of multi-word units that are infrequent and non-compositional, whereas lexical bundles — phrases like *I think that* or *at the end of* — are both frequent and compositional. More frequent and less compositional strings have been found to be processed faster. Some argue that this shows that we store those strings as wholes (Beckner et al., 2009; Conklin and Schmitt, 2012; Pawley and Syder, 1983).

1.2 The Why

The idea that we store units larger than a word is uncontroversial. *It takes two to tango* cannot be understood by simply combining the meanings of its single words — you have to rely on the stored meaning of the whole. Likewise, infrequent and fully compositional word combinations are most likely composed and parsed on-line. However, controversy arises the further we move into the lower right corner of Figure 1.1.

Dual-system theories of language assume that language consists of a grammar and a lexicon (Pinker and Ullman, 2002). The lexicon contains all that cannot be computed; the grammar contains rules that are used on all that is stored to compute new forms. As such, the lexicon does not contain any redun-

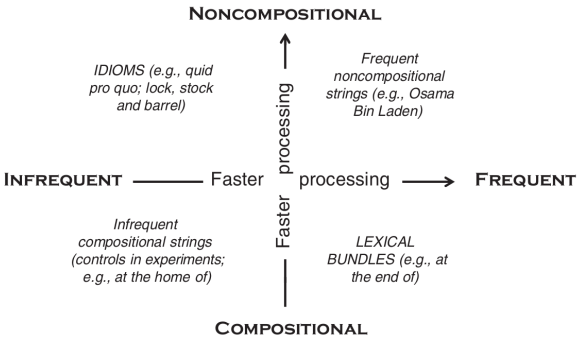

Figure 1.1: A typology of multi-word units. This thesis focuses on the units from the lower right corner, lexical bundles. Figure taken from Wray (2012, p. 241).

dancy: If the grammar can compute something, then the end product of that computation will not be stored in the lexicon. The prediction is, then, that compositional multi-word units are not stored or used as wholes in processing.

Single-system theories, on the other hand, argue that there is no principled difference between word forms and grammar in processing. They propose that both rules and words are part of the same system, where the end products of applying regular rules to words can be stored. Redundancy is assumed to be prevalent (Dąbrowska, 2014; Snider and Arnon, 2012). In such a system, it is possible that frequent and compositional multi-word units, lexical bundles, are redundantly stored.

The fact that storage is possible does not provide an explanation of why such units would be stored, and by what mechanisms. Researchers have argued that we store redundant forms in our long-term memory to compensate for our limited working memory capacities. Having to repeatedly combine single words will take up more working memory resources than storing and later retrieving a multi-word unit, which in turn makes language processing faster and more efficient (Conklin and Schmitt, 2012; Pawley and Syder, 1983).

Usage based-linguistics (Bybee, 2006, 2010; Green, 2017; Goldberg, 2003; Tomasello, 2009) has proposed the cognitive mechanism by which lexical bundles come into being: chunking. Through repeated exposure frequent combinations can become more fixed and merge together into a holistic unit. Chunking is seen in all kinds of cognitive domains, from action sequences such as cycling

or tying your shoe laces, to remembering the notes that make up a melody (Bybee, 2010).

Many researchers are hesitant to argue that frequent multi-word units are stored holistically (Arnon and Cohen Priva, 2013; Siyanova-Chanturia, 2015). There is a growing body of experimental work showing that lexical bundles are read, understood, and pronounced faster than their infrequent matched controls (Arnon and Snider, 2010; Bannard and Matthews, 2008; Tremblay et al., 2012). Even so, faster processing does not necessarily entail holistic storage. Phrasal frequency effects could reflect experience and therefore greater proficiency in combining and decomposing those specific combinations (Tremblay et al., 2011). Individual words still play a role in processing in lexical bundles (Arnon and Snider, 2010; Siyanova-Chanturia, 2015), and experiments have shown that even single words can prime idioms or other non-compositional phrases, testifying to the existence of internal structure and against the notion of holistic blocks (Sprenger et al., 2006).

1.3 The How

When studying a specific phenomenon, it is important to employ different methods. Each method might shed a different light on that phenomenon, so that one learns from considering where the insights converge and diverge (Rayner, 1998). Researchers have used different paradigms and experimental methods to investigate the processing of different types of multi-word units. Most research has focused on either reading or speaking, while listening to multi-word units has received far less attention. In the following, I will focus on lexical bundle processing by discussing what we know so far about reading, listening to, and producing lexical bundles.

1.3.1 Reading

There are several ways to study the processing of lexical bundles with reading paradigms. By using simple behavioral methods, such as lexical decision tasks or self-paced reading, researchers have found that lexical bundles are processed faster than matched control sequences. Furthermore, by using eye-tracking, researchers have learned more about the time course of processing.

Durrant and Doherty (2010) used a lexical decision task to see whether the first word of a frequent collocation, such as *mental*, would prime its second word, here *picture*. When the prime was masked, however, they only found a significant priming effect when the two words of the collocation were also associates of each other, as in the collocation *card game*. As such they only found convincing evidence for associative priming. Perhaps the priming paradigm was not sensitive enough to detect any phrasal effects, or presenting only one word at a time did not prompt any lexical bundle processing.

Jiang and Nekrasova (2007) as well as Arnon and Snider (2010) took the lexical decision task a bit further and conducted a phrasal-decision task, where participants were asked to judge if phrases were grammatical strings or not. Matched pairs of lexical bundles and control phrases were used. One word from the lexical bundle was replaced with a word similar in length and frequency to create a matched control phrase, such that the pairs only differed in their phrasal frequencies. Both studies found faster reaction times for lexical bundles than control phrases. Jiang and Nekrasova (2007) also identified phrasal frequency effects in proficient non-native speakers, while Arnon and Snider (2010) noted that the effects could be observed across the whole frequency range, with low- and mid-frequency phrases also being processed faster than their matched controls.

Providing grammaticality judgments is not a very natural task — it involves making meta-linguistic decisions and might not be the best reflection of what language users do in daily life. A more naturalistic task is self-paced reading, where people read through whole sentences, or even paragraphs, piece by piece. Tremblay et al. (2009) found that sentences containing lexical bundles are read faster, but only if these sentences are presented chunk-by-chunk or as a whole. Word-by-word presentation seems to disrupt phrasal frequency effects — which might explain the absence of collocational priming effects in the study conducted by Durrant and Doherty (2010).

Pressing a button before one can move on with reading is still not very similar to the way we normally read. Also, it yields only one measure: The latencies of button presses. Eye-tracking, on the other hand, generates many different measures, which reflect how difficult processing is, and how processing proceeds over time. For example, the harder it is to process a text, the longer a duration will last, and the more fixations a reader will need. Moreover, looking at differences between early and later fixations tells something about how processing proceeds over time (Rayner, 1998). Because one can study the effects of both single words and phrases at the same time, eye-tracking offers exciting new insights into the processing of multi-word units (Siyanova-Chanturia, 2013).

The first study looking into the eye-movements of people reading multi-word units is Underwood et al. (2004). The authors compared the number of fixations and their durations on the final words of idioms and novel phrases. Identical lexical items attracted fewer and shorter fixations in idioms than in non-idiomatic phrases. This was interpreted as reflecting holistic storage and processing of idioms. Similar results were found by Siyanova-Chanturia et al. (2011a), who found that readers need fewer and shorter fixations for idiomatic than non-idiomatic phrases, and that these phrases require less re-reading and re-analysis.

Moving on to non-idiomatic multi-word units, Siyanova-Chanturia et al. (2011b) studied the eye-movements of people reading binominal phrases such as *bride and groom*. These types of phrases are similar to lexical bundles in that they are compositional, but they are not as frequent. They lie somewhere in the

bottom middle of Wray's model as presented in Figure 1.1. Siyanova-Chanturia et al. (2011b) compared these phrases with their reversed counterparts (i.e. *groom and bride*), which are identical in meaning and single-word frequency, but different in phrasal frequency. Both early measures (first pass reading time) and late measures (total reading time and fixation count) were influenced by phrasal frequency, where more frequent phrases were read faster, with fewer fixations, than less frequent phrases.

Tremblay and Baayen (2010) looked at lexical bundles. They employed an immediate recall task, where participants were first shown six four-word sequences, and then asked to type in as many sequences as they could remember. During the presentation of the sequences, the authors collected EEG data. They found that both single words and sequence-internal trigrams modulated the behavioral results, indicating that both parts and wholes are used in processing. Furthermore, phrasal frequencies modulated the electrophysiological signal from very early on — roughly 100 ms after presentation onset. This suggests that the whole string must be accessed and retrieved as a holistic chunk, as there is no way that single words can be retrieved and combined in such a short time frame.

Miwa et al. (2017) also found early effects of holistic processing — in this case frequency effects of Japanese trimorphemic compounds. In a lexical decision experiment coupled with eye-tracking, the first fixation durations were modulated by the full compound frequency. Importantly, the frequencies of the single morpheme also played a role in processing.

Overall, researchers have found that higher phrasal frequencies correlate with shorter reading times, and that frequent multi-word units need fewer fixations and re-analysis. Moreover, both single words and the multi-word unit play a role in processing, casting doubt on the idea that multi-word units are stored as unanalyzed chunks.

All of these studies focus on younger adults. However, if language experience is indeed the driving force behind the emergence of lexical, as usage-based theories of language propose, then more language experience should lead to differences in the representation or processing of lexical bundles between younger and older adults. This brings us to the first research question of this thesis: **How do adults read lexical bundles, and are there differences in reading behavior between younger and older adults?** Chapter 2 presents an eye-tracking study of both younger and older adults reading lexical bundles.

Previous work has emphasized the role of traditional frequency measures in processing; but **Which factors other than frequency play a role in reading lexical bundles?**. Moreover, knowing which factors play a role in processing does not answer the questions **How does lexical access to lexical bundles proceed?**, and **What is their status in the lexicon?**. Therefore, Chapter 4 presents a computational model of lexical bundles and discusses how this model can shed further light on how lexical bundles are read and accessed and how this, in turn, sheds light on their very nature.

1.3.2 Listening

While there is a lot of research on reading multi-word units, it is not yet completely clear what happens when people listen to them. In Sosa and MacFarlane (2002)'s study, people listened to utterances containing collocations with the word *of*. People were asked to press a button as soon as they heard *of*. The more frequent the collocation, the slower people were, and the more misses they made. According to the authors, this indicates that frequent collocations are stored holistically. Because of their holistic form, people do not automatically deconstruct the collocation into its constituent parts, and therefore cannot detect single words immediately, leading to slower response latencies.

In a sentence recall task, Tremblay et al. (2011) presented participants with spoken sentences containing lexical bundles. They found that recall rate correlated positively with phrasal frequencies, such that sentences containing more common lexical bundles were remembered correctly more often. Tremblay et al. (2011) conclude that lexical bundles are a relevant unit in processing.

To summarize, we have evidence that people are better at recalling lexical bundles that they have listened to, and that during listening, they do not always seem to parse the single words contained in these lexical bundles. To add new research to the issue of listening to lexical bundles, Chapter 3 investigates listening to frequent lexical bundles and infrequent matched controls to answer the research questions **Is there a difference in electrophysiological brain responses when listening to frequent lexical bundles and infrequent matched controls?**, **Which factors influence the electrophysiological brain response when listening to lexical bundles?**, and **What is the time course of processing of auditorily presented lexical bundles?**.

1.3.3 Speaking

When speaking, we occasionally make slips of the tongue. These slips may involve phonemes, clusters of phonemes, syllables, morphemes, words, or even parts of phrases. Slips are claimed to only occur within linguistic units — they do not involve random exchanges of phonemes across a unit boundary (Kuiper et al., 2007). Because slips are found in multi-word units, it seems likely that these units have a separate entry in the mental lexicon.

To test if children make use of lexical bundles, as proposed by usage-based theories (e.g. Tomasello, 2009), Bannard and Matthews (2008) used a sentence-repetition test with 2- and 3-year-old children to see how well they could repeat different utterances. These utterances were taken from a corpus containing child-directed speech, and consisted of frequent phrases such as *sit in your chair*, and were matched to infrequent phrases such as *sit in your truck*. Both groups of children were more likely to repeat the frequent lexical bundles and made fewer mistakes when doing so. For the 3-year-olds it was even the case that the durations of their productions were significantly modulated by phrasal frequency, with higher frequencies correlating with faster productions.

Several studies have looked at how adults process lexical bundles. Tremblay and Tucker (2011) had people read out loud four-word strings from a computer screen, and they measured the onsets and durations of those utterances. The authors found effects of unigram, bigram, trigram, and quadgram frequencies. This suggests parallel processing of both the whole lexical bundle and its constituent parts.

To advance insights into the production of lexical bundles, Arnon and Cohen Priva (2013) looked at both experimentally elicited speech and naturalistic speech taken from a corpus. They also tested whether a lexical bundle has to be a single syntactic constituent in order for it to show phrasal frequency effects. In accordance with other findings, higher phrasal frequencies correlated with shorter durations. Notably, these effects occurred both within and across syntactic boundaries.

In a follow-up study, Arnon and Priva (2014) observed that phrasal frequencies play a role across the whole frequency range. Lower phrasal frequencies lead to a higher prominence of the effects of single words, whereas higher phrasal frequencies lead to a reduced prominence of single word frequencies. Crucially, the effect of single word frequencies does not disappear, showing that the storage and processing of frequent multi-word units does not necessarily involve any holistic, unanalyzable blocks of language. The parallel effects of single-word and multi-word unit frequencies are similar to the findings of Tremblay and Tucker (2011).

Besides reading words to elicit multi-word unit production, researchers have also used picture-naming paradigms. Using Spanish multi-word units, Janssen and Barber (2012) presented participants with colored and superimposed line drawings to elicit noun + adjective, noun + noun and determiner + noun + adjective structures. Naming latencies decreased with increasing phrasal frequencies, suggesting a role for multi-word units in production.

However, Hendrix et al. (2017) did not find any effects of phrasal frequencies in the naming latencies of nouns in frequent prepositional phrases. They did, however, find qualitatively different patterns for word frequencies and phrasal frequencies in the ERP signal: Word frequencies were characterized by oscillations in the lower theta range, whereas phrasal frequencies did not elicit any theta oscillations, but showed a prolonged negativity for multi-word units with higher phrasal frequencies.

In short, the evidence from production studies shows that higher phrasal frequencies lead to shorter production latencies and better recall. Importantly, many studies have shown that both single words and multi-word units affect production.

Building on these existing studies, the second part of Chapter 4 consists of a production study where participants read high-frequency lexical bundles out loud from a computer screen. By employing measures extracted from a computational model incorporating those lexical bundles to model that data, I aim to answer the question **Are there other factors over and above traditional frequency measures that play a role in reading out loud**

high-frequency lexical bundles?.

1.4 Quantifying processing

Experimental measures taken from either eye-tracking, EEG, or production studies are but a pale reflection of what is really happening during processing. Processing language is an intricate, multi-faceted process, and it is therefore crucial to try to quantify its subprocesses. This can for example be done by fitting a statistical model on some dependent variable. The predictors of such a model will consist of the factors that the experimenter has under her control, such as the frequencies of lexical bundles and their subparts, as well as the factors that are outside her control, such as the participants' mental state during the experiment. Another way is by building a computational model, which forces the researcher to explicitly specify certain aspects of processing and lexical bundles so as to obtain predictions on how lexical bundles will behave in processing.

1.4.1 Statistical modeling

We live in an exciting time where statistical modeling, machine learning algorithms and computational power are constantly improving. Moreover, these methods are increasingly accessible to a wider group of researchers due to easy-to-use implementations in software such as the statistical programming language R (R Core Team, 2017).

In this thesis, a large part of understanding the processing of lexical bundles comes from applying advanced statistical models to experimental data. Because previous research has shown that both parts and wholes of lexical bundles simultaneously play a role in processing, it is important to use techniques that can take all these factors into account, while at the same time trying to account for all the unknown noise that affects experiments: How much coffee did a participant drink today? Did he sleep well? Does she have experience with these types of experiments? Does the noise from the construction workers distract the participant? Does this lexical bundle have an unexpected effect on the participant because she just read a newspaper article containing that very unit? In what follows, I will briefly present the key models used in this thesis.

Generalized Additive Mixed-Effects Models

Nature is full of dynamic and nonlinear systems — language is but one of them. We cannot assume therefore that all experimental data are linear: We need to take into account nonlinearity. Generalized additive mixed-effects models (GAMMs, Hastie and Tibshirani, 1990; Wood, 2006) are regression models that can model nonlinear relations in the data. This is done by means of so-called spline-based smoother functions, which are functions that model a nonlinear

(or so-called 'wiggly') curve on the relation of a predictor and the outcome variable of interest.

In order to fit a reliable regression model, all data points need to be independent from each other. However, this is never the case with data gathered in psycholinguistic experiments: The data points produced by one person are always correlated to each other because that person has some unique characteristics that will affect in a certain way all responses he gives. Other participants will have other unique characteristics, which in turn affect their responses in other, unique, ways. The same is true for experimental items: Each item might have certain characteristics that are not or cannot be explicitly included in the statistical model, but that do introduce commonalities into the responses of all participants to that specific item. For example, imagine a situation where there are a lot of news items about an alpaca who was left behind in a city center[1]. In that situation, the word 'alpaca' will be much more salient to participants than it normally would be, leading to commonalities in how these participants will react to that specific word.

GAMMs incorporate random-effects structures that take this non-random noise into account. The random-effects part of a model introduces parameters specifically for these commonalities in responses from individuals to individual items. This makes the other model parameters more accurate (Baayen et al., 2008; Barr et al., 2013; Bates et al., 2015).

Besides offering the possibility to model nonlinear relations and allowing for a random-effects structure, GAMMs can also include predictors that model the time course of the whole experiment. This is essential as each participant will have a different attentional flow throughout the experiment: Some participants might be alert in the beginning, responding quickly, but then losing attention along the way, thereby responding more slowly. Other participants start off slowly, and get faster over the course of the experiment (Baayen et al., 2017a). Entering a predictor that describes this behavior over time will also improve the model fit and allow for better estimates of the predictors of interest.

However, regression modeling has several disadvantages. It is intolerant to multicollinearity: When two or more predictors are highly correlated, their model parameters can no longer be trusted (Wurm and Fisicaro, 2014). A model parameter is the estimation of the shape, size and direction of an effect — in other words, crucial information in understanding what is happening in the data and for testing if hypotheses hold. Multicollinearity is especially problematic when modeling behavioral and neural responses to lexical bundle frequencies, as the frequency of the whole lexical bundle is very often highly correlated to the frequencies of its constituent n-grams (e.g. single words, bigrams, trigrams).

Researchers have resorted to different techniques to deal with multicollinearity. One is reducing the number of dimensions, by creating a composite variable from the correlated predictors (for example by Principal Component Analysis

[1]See for example this news item for more information on alpaca Teddy (in Dutch). https://www.rtlnieuws.nl/nederland/bert-helpt-gedumpte-haarlemse-alpaca-geen-dier-kun-je-zo-behandelen

(Baayen, 2008)). Another technique is residualization, where a variable A is regressed on a collinear variable B. The residuals of that regression are then entered into the model — the idea being that these residuals are what is 'left' of variable A when one takes out the parts that correlate with B (see for example Tremblay and Tucker, 2011, for an example of how to deal with a large number of correlated predictors). However, these techniques also have their disadvantages. It is not so straightforward to understand what a composite or a residualized variable is, which makes a model with these types of predictors hard to interpret. Moreover, residualization is not a remedy for multicollinearity (Wurm and Fisicaro, 2014).

Other problems with mixed-effects regression analyses are 1) the need for normally distributed data (as all experimental linguists know, no data are ever normally distributed); 2) the question of how to specify the random-effects structure of the model (Barr et al., 2013; Bates et al., 2015); 3) the biases of the researcher in choosing which interactions to enter into the model, thereby potentially overseeing important interactions; and 4) the fact that forward and backward model fitting is notoriously susceptible to the order in which predictors and interactions are added and deleted (Strobl et al., 2009).

Despite its disadvantages, regression modeling, and especially mixed-effects modeling, has proved to be a very useful tool in modeling experimental data, and has enhanced our understanding of the intricate processes used in language production and comprehension. One certainly has to keep in mind its shortcomings, but the advantage of having a model that allows for non-linearity, combined with the power of random-effects, make GAMMs the statistical model of choice for the data in Chapters 2 and 4.

Conditional Inference Random Forest Models

A way to avoid the problems commonly encountered in regression modeling, is by using non-parametric methods that do not require the researcher to specify in which order predictors need to be added or deleted, and which interactions should be tested. We are all but humans — machines are better suited to do these tasks for us.

A popular machine learning technique, Conditional Inference Random Forest Models (CForests), does not suffer from the drawbacks discussed in the previous section. As these are non-parametric models, the data need not be normally distributed. Furthermore, these models are very robust to noise, and a large part of the modeling process is data-driven, instead of being based on fallible human decisions. This way, unexpected or complex higher-order interactions present in the data will still be taken into account, even if the human modeler never thought of including them.

A random forest consists of a large set of randomly built decision trees. Consider Figure 1.4.1 for an example of a decision tree. In this decision tree, a model is presented that classifies animals into two categories: Cats and dogs. The model uses a continuous variable (body weight) and binary variables (does

it attack a laser, and does it love you?) to predict which category is most likely. If you encounter an animal that weights two kilos, and who does not attack any lasers, then you will most likely have a dog in front of you (probably a chihuahua). Note however, that the categories at the bottom of the figure only show the most likely candidates — it is still possible, although less likely, that the small creature that does not care about the laser is a tired cat.

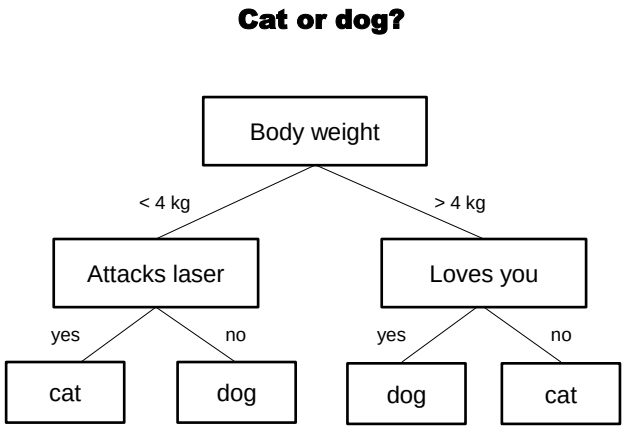

Figure 1.2: A decision tree that helps to distinguish between human's favorite pets, cats and dogs.

As has become clear from the example above, decision trees are models that use a tree-like graph to visualize the internal structure of data. Each node represents an attribute (predictor) on which a decision (split) of the data is based. The leaves or end nodes of the tree represent the different groups that the model identified in the data, and shows the predicted outcomes. This could be the result of a coin flip (heads or tails), the predicted reaction latency in a production experiment (the subgroup of responses to content words with a length of 6 letters or more is on average 720 ms), or the predicted voltage of a participant hearing a lexical bundle (the subgroup of lexical bundles that has more than 12 letters and whose first bigram has a higher frequency than 1,000, correlates with an average signal of -1.2 microVolt).

A decision tree is constructed by repeatedly splitting the data into two, based on whichever predictor does the best job at identifying two subsets in the data. In our pet example in Figure 1.2. the body weight has been selected as the best predictor at dividing pets into cats and dogs. After this first split, it turned out that for animals less than 4 kilos, the predictor 'attacks laser' is best at dividing cats and dogs, whereas the predictor 'loves you' plays the largest role in splitting the data for animals that weigh more than four kilos. This process of binary splits continues until, for example, none of the predictors

reaches significance in a certain subset (Hothorn et al., 2006; Strobl et al., 2009). The resulting tree will contain information on which predictors are important, interactions within the data, and the number of data points that fall into each subset.

The decision trees used for CForest modeling incorporate another feature: *variable preselection*. Instead of always testing all predictors on the data and on all of its subsets, a subset of the predictors is randomly selected for each split within the tree. That way, even weaker predictors with small and subtle effects, that otherwise might have gone unnoticed, have a greater chance of entering the model. Variable preselection results in a diverse set of trees that form the forest. By aggregating over these trees, even subtle effects and potentially informative but unexpected interactions are likely to surface.

To make the set of trees even more diverse and therefore more stable to noise (Strobl et al., 2009), *bagging* can also be applied to the data. Bagging means that every tree in the forest is grown on a random subset of the data. By using variable preselection and bagging, the results of the CForest modeling in Chapter 3 are very robust, precise and contain information that other types of modeling might have never brought to light.

Random forests, and specifically CForests, have been applied in diverse fields such as genetics, epidemiology, medicine, and lately also in psychological and linguistics datasets (Tagliamonte and Baayen, 2012; McWhinney et al., 2016). As CForests are able to model all kinds of functional relations between predictors and an outcome variable over time, they are well-suited to handling many collinear predictors, such as frequency values of trigrams and their constituent bigrams and single words.

1.4.2 Computational modeling

To further understand what a lexical bundle is, it is helpful to explicitly model the processing of lexical bundles in a computational model. This way, one can test if and how the model's predictions fit experimental data, and if they do, study how the model functions. The model of choice in this dissertation is a Naive Discriminative Learning or NDL model (Baayen et al., 2011; Baayen and Ramscar, 2015).

Naive Discriminative Learning (NDL)

NDL is a theory of lexical processing, which is made explicit in a computational model. The training phase of the model can be seen as an L1 acquisition process, whereas the stable end state of the model, where it has reached an equilibrium, can be considered as the adult state of the linguistic system of the learner. This end state of the model provides a mathematical characterization of the state of the lexicon, and can be used to derive several features that describe on-line processing. Interestingly, these features have proven to be excellent predictors of a wide range of linguistic phenomena such as lexical decision latencies, word

frequency effects, phrasal frequency effects, and ERP amplitudes. Moreover, predictions following from NDL models are consistent with the performance of young infants in an auditory comprehension task (Baayen et al., 2011; Baayen and Ramscar, 2015).

An NDL model features a simple two-layer network where input units, such as sounds or written letters, form the cues that are connected to a set of outcomes. These outcomes consist of *lexomes*, which are pointers to a location in semantic space. See Figure 1.4.2 for an example of a small NDL network. It is important to note that lexomes are neither form nor meaning, but stable mediators between variable linguistic forms and meanings (Milin et al., 2017; Baayen et al., 2017b). Because an NDL network has no hidden layers, the way its connections are formed over time is a relatively straightforward process.

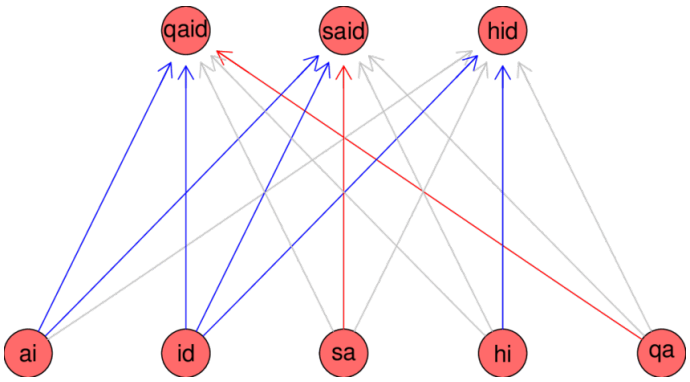

Figure 1.3: An NDL model with five digraphs as cues, and three lexemes as outcomes. Figure taken from Baayen and Ramscar (2015).

In an NDL network, all cues and outcomes are connected to each other. The weights of these connections are established by training the model on a large set of sentences. First, the form of both the cues and the outcomes have to be defined: Often the cues are formed by single letters or combinations of letters (Baayen et al., 2017b, 2016a), and the outcomes are pointers to the meanings of single words. However, outcomes can also point to grammatical features (Baayen et al., 2011), idioms (Geeraert et al., 2017), or, as in Chapter 4, lexical bundles.

In the training phase, the model goes over a large set of sentences one by one, and at each sentence, updates its connections weights between cues and outcomes. The Rescorla-Wagner learning rule (Rescorla et al., 1972) specifies how these connection weights are updated. Rescorla-Wagner equations have been quite successful at describing how animals and humans learn (Arnon and Ramscar, 2012), which motivates their use in a model that tries to capture how humans build up their linguistic knowledge over time.

The way the Rescorla-Wagner equations work is by comparing the predic-

tions made on the basis of the input cues (i.e. what outcomes are expected given these letters?) and the actual outcomes. When a prediction is correct, the association weight between the cue and outcome is strengthened. Conversely, when a prediction is incorrect, that is, when a cue occurs without an outcome, their association weight is weakened.

A cue is informative and thus discriminative if strong connection weights lead to only a small number of outcomes. However, if a cue is more or less evenly connected to a lot of different outcomes, then this specific cue is not a strong predictor of any outcome. Determiners are bad predictors of the identity of any multi-word unit, whereas the word *happily* is a strong discriminative cue for the outcome *happily ever after*.

After the training phase, the weights of the model provide a mathematical characterization of the state of the lexicon. More specifically, they indicate how well outcomes can be discriminated given a certain set of input cues. From this network predictions can be made: One can extract features that can subsequently be put into a statistical model that aims to describe experimental data.

Implementing lexical bundles in NDL

Not only is it relatively easy to understand the inner workings of a NDL model, its way of learning over time is implemented using a cognitively plausible learning algorithm. This algorithm has been proven useful in describing (implicit) learning in animals and humans (Ramscar et al., 2010, 2013; Ramscar and Yarlett, 2007), thereby positioning this model of linguistic behavior also in an evolutionarily and cognitively plausible context. In order to further understand the processing of lexical bundles, it is therefore worthwhile to see how well an NDL model that incorporates lexical bundles performs in explaining experimental data.

Implementing lexical bundles in NDL would amount to implementing these units as lexomes — in other words, items that function as unitary items in processing. Not only will this implementation shed more light on lexical bundle processing and the factors that play a role therein, but it will also provide a characterization of lexical bundles, when considering their status in the model.

1.5 This dissertation

Chapter 2 comprises a study on reading lexical bundles by both younger and older adults. Their eye movements were monitored with an eye-tracker, and these data have been analyzed using GAMMs. Results showed no differences in the processing of lexical bundles, but did show differences between the age groups in how they processed single words and bigrams. These differences are argued to originate from changes in cognitive and physical skills.

Chapter 3 is about listening to lexical bundles. While being hooked up to an EEG machine, participants listened to a diverse set of lexical bundles. CForest modeling of the results reveals the time course of lexical bundle processing and the intricate roles that single word frequencies, bigram frequencies, and trigram frequencies play.

Chapter 4 combines two behavioral experiments, where people read and produced lexical bundles, with an NDL model containing these same lexical bundles. Predictors taken from the NDL model are quite successful at capturing variance in the experimental data, testifying to the usefulness of using features other than traditional frequency measures to explain lexical bundle processing.

The last chapter, Chapter 5, discusses the converging and diverging results coming from these different experimental techniques and statistical modeling and aims to provide a rich and multi-faceted overview of lexical bundle processing.

CHAPTER 2

Reading

Old and young: How language experiences (do not) shape the reading of lexical bundles

Saskia E. Lensink, Niels O. Schiller, Arie Verhagen

abstract
Repeated exposure to common combinations of words is believed to result in the use of these combinations as chunks in language processing. If this is true, then it is expected that older participants will process lexical bundles differently due to their larger experience with language. We report on an eye-tracking study where a group of younger adults and a group of older adults read through a list of frequent Dutch lexical bundles. We made use of non-linear mixed-effects statistical models to study which linguistic and non-linguistic features play a role in reading, and what functional shape their relationship with different dependent reading measures has. We found no age-related differences in how lexical bundles are processed. The results show an intricate interplay of both language-related factors such as frequencies and perceptual/oculomotor factors such as viewing position and regressions. We furthermore found an Inverted Frequency Effect for trigrams, where more trigrams elicit longer fixation durations. We argue that this effect might be caused by either lexical competition, phrasal processing that is different from word processing, or a reading strategy.

Keywords eye-tracking; aging; multi-word units; mixed-effects modeling

2.1 Introduction

Usage-based theories of language propose that language usage shapes language representations (Bybee, 2010; Goldberg, 2003). These theories predict that if certain words co-occur often in usage, they become chunked over time. These multi-word units are thought to increase the speed and efficiency of processing. Indeed, recent experimental evidence has shown that language users employ multi-word units in processing. They speed up production latencies, modulate electrophysiological brain responses, and influence reading behavior (Arnon and Snider, 2010; Hendrix et al., 2017; Lensink et al., submitted; Siyanova-Chanturia, 2013; Tremblay and Tucker, 2011). However, most research focuses on relatively young adult participants.

If usage indeed shapes the way language is represented, then accumulated years of language experience will have implications for the role multi-word units play in storage and processing. From this usage-based perspective two inconsistent predictions can be derived: Either older adults employ more multi-word units actively to speed up processing, or they employ less multi-word

units, but instead are more efficient in parsing sentences than younger readers. This study sets out to explore which of these two possibilities applies to multi-word unit processing.

Accumulated experience with multi-word units could lead to a stronger representation of those units, and, due to a growing experience with other and more diverse forms, older adults would have a larger set of multi-word units in general. This could lead to a greater use of and reliance on multi-word units in processing. Conversely, older adults could employ less or no multi-word units in processing if their accumulated experience with language improves skills in combining single words and parsing sentences. After all, it is well-known that children at first use unanalyzed chunks in their speech productions, which they start to analyze and decompose only later (Tomasello, 2009).

To explore if and how the use of multi-word units in processing changes over the years, we ran an eye-tracking experiment where we visually presented frequent lexical bundles to two groups: People in their twenties and people in their sixties. In what follows, we will first discuss what general differences exist between younger and older adults in language processing. We will zoom in and discuss what is known about the processing of multi-word units in reading paradigms, and discuss our predictions on how multi-word unit processing will differ between younger and older readers.

2.1.1 Language processing in younger and older adults

Research on the influence of aging on language processing has shown that production skills deteriorate over the years, whereas most comprehension skills remain intact (Shafto and Tyler, 2014). Vocabulary size even grows over the years (Keuleers et al., 2015). Also, there is no evidence for neural reorganization of core language areas due to age (Shafto and Tyler, 2014). So even though aging is often associated with becoming slower, less precise, and more forgetful, most linguistic abilities stay stable or improve over the years. However, when it comes to reading behavior, there are several differences between younger and older adults.

Generally, there are more quantitative than qualitative differences (Laubrock et al., 2006). Older readers have a smaller and a more symmetrical perceptual span (Rayner et al., 2009). They are affected more by the blocking of foveal information than younger readers, indicating that they are less efficient in processing parafoveal information (Rayner et al., 2014). Furthermore, seniors make more and longer fixations, tend to skip more words, make longer saccades, and more regressions, resulting in slower reading overall (Laubrock et al., 2006; Rayner et al., 2006, 2009).

Therefore, age-related differences in reading lexical bundles are to be expected. The next section will outline what we know about reading lexical bundles by younger adults.

2.1.2 Reading lexical bundles in younger and older adults

In this study, we use eye-tracking to study if and how multi-word units are processed differently by younger and older adults. Eye-tracking offers detailed insights into how cognition handles written text, tracking different dependent measures over time: First fixations reflect early processes, while later fixations reflect later processes, and the number of fixations reflect the overall processing difficulty over the whole region considered. Early processes of reading involve lexical access and early integration of information; later processes involve further integration of that information, re-analysis, and recovery from processing difficulties (Siyanova-Chanturia, 2013). There are several advantages to using eye-tracking when studying the processing of lexical bundles. Not only can processing be tracked precisely over time, it is also possible to study both phrase- and word level patterns at the same time (Carrol and Conklin, 2015).

In an eye-tracking study, Siyanova-Chanturia et al. (2011b) investigated how frequent binominal phrases such as *bride and groom* were processed in contrast to their reversed counterparts, *groom and bride*. The authors found clear phrasal frequency effects, where more frequent configurations such as *bride and groom* were read faster than their reversed counterparts. Crucially, only phrasal frequencies surfaced as significant predictors; the frequencies of the first and second content word did not influence reading times. A comparable effect was found in both early and late measures.

Looking at a wider range of compositional multi-word units, Lensink et al. (submitted) studied the reading of frequent Dutch lexical bundles. They used predictors from a naive discriminative model (Baayen et al., 2011) to model reading measures. They found that participants are already sensitive to properties of a lexical bundle from the first fixation onwards. Research on the reading of compounds has also shown early full-form frequency effects, sometimes as early as in the first fixation (Kuperman et al., 2009; Miwa et al., 2017).

2.1.3 The present study

When considering previous research, it is not immediately clear if and how the reading of multi-word units will differ between younger and older adults. However, it is likely that the general reading patterns differ quantitatively, with older adults using more and longer fixations, and, due to their shorter perceptual span (Rayner et al., 2009), we expect older adults to react differently to shorter and longer trigrams.

It is unclear if the two groups will respond differently to single word frequencies, bigram frequencies, or trigram frequencies, and what the direction of the effects might be. Seniors could have more and stronger representations of lexical bundles due to a larger and more diverse experience with language. In that case, stronger facilitative effects of frequent bigrams and trigrams are expected in older adults, leading to shorter fixation durations and less fixations than in younger adults.

Alternatively, older adults might use fewer multi-word units in processing due to their larger experience and greater efficiency with retrieving and combining single words. In that case, they will make less or no use of multi-word units. This would manifest in the data as a weak or no phrasal frequency effect for older adults.

This study explores these possibilities by collecting eye-tracking data on reading lexical bundles from both older and younger adults, and by statistically modeling these data[1]. Section 2.2 reports on the eye-tracking experiments conducted on both groups. In Section 2.3 we report on the results of these experiments. Our interpretations, the limitations, and wider implications of this study are discussed in Section 2.4.

2.2 Materials and Methods

2.2.1 Participants

We recruited two groups of participants: A group of people between 18 and 30 years of age, and a group of people between 60 and 72 years of age. We will refer to these groups as the 'younger' and the 'older' group, respectively. The younger group consisted of 32 participants, from which we lost the data of seven participants due to technical issues. The remaining 25 (10 male) were on average 21.4 years old. For the older group, we recruited 31 participants, out of which we lost the data of four participants due to technical issues. The remaining 27 (15 male) were on average 66.2 years old. All participants were native speakers of Dutch, had normal or corrected-to-normal vision, and received a monetary or culinary[2] reward for their time.

2.2.2 Stimuli

A set of three hundred trigrams was randomly extracted from the top one per cent of the most frequent trigrams in the Dutch OpenSoNaR corpus (Oostdijk et al., 2013). We sampled from the top one per cent to ensure that under any usage-based account, these combinations are predicted to be stored as chunks and are thus likely to influence processing over and above the single words they consist of (Bybee, 2010). We only included transparant trigrams where the meaning of the whole can be largely deduced from the meanings of its parts, so-called 'lexical bundles' (Tremblay and Tucker, 2011; Wray, 2012).

The stimuli were put in two different experimental lists. We ensured that there was no semantic or phonological overlap between stimuli in two consecutive trials, to prevent priming effects.

[1]The data and analyses that support the findings of this study are openly available in Figshare at https://doi.org/10.6084/m9.figshare.5982340.v1.

[2]Most participants from the older group refused to accept the monetary reward, so the experimenter decided to offer them a cup of coffee or tea with a snack instead.

Trigram	Continuation	Correct
de hele wereld 'the whole world'	*staat de kast* 'stands the closet'	incorrect
na de pauze 'after the break'	*loop ik terug* 'I walk back'	correct
ik denk dat 'I think that'	*zij druk is* 'she is busy'	correct

Table 2.1: Three trials containing both a trigram and a follow-up sentence fragment, of which the participants had to indicate whether or not it constituted a grammatical continuation of the previous trigram.

2.2.3 Procedure

Prior to the experiments, participants filled in a questionnaire about their language background and gave informed, written consent. They were seated in a sound-proof room and were asked to put their chin on a head rest to minimize head movements. Their dominant eye was recorded with an Eyelink 1000 eye-tracker (SR Research Ltd) at a 500 Hz sampling rate. At first, the experimenter performed eye calibration using a 9-point calibration procedure.

After successful calibration, participants saw a screen with written instructions. The experiment started with a practice block of five trials, where participants were asked to read silently through a set of trigrams presented one by one on the computer screen. After the practice block, participants continued reading the full stimulus set of three hundred trigrams. At random intervals, a trigram was followed by a small fragment, which could either form a grammatical or an ungrammatical continuation of the previous trigram. This continuation stayed on the screen until the participant clicked on a box with 'correct' or 'incorrect' with a computer mouse. See Table 2.1 for some example trials. Participants received immediate feedback by means of a screen displaying either the word 'correct' or 'incorrect'.

The experiments consisted of three blocks, with short breaks in between. At the start of each trial, a fixation point was presented for 500 ms at a fixed position at the left-hand side of the screen, to ensure reading from left to right. Trigrams were presented in a black, mono-spaced font (Consolas, size 22) against a white background for 1,200 ms. Trials were separated by an inter-stimulus interval of 1,000 ms.

2.2.4 Generalized Additive Mixed-Effects models

To model the differences in reading behavior of the two groups, we used generalized additive mixed-effects models (GAMMs) (Hastie and Tibshirani, 1990; Wood, 2006). GAMMs are able to model nonlinear relations between predictors and an outcome variable using spline-based smoother functions, creating wiggly curves for single predictors and wiggly (hyper-)surfaces for interactions between predictors. Moreover, they incorporate random effects, that can model individual variation either as variations in the intercepts or the slopes, or as

a combination thereof, as variations in wiggly curves. To prevent overfitting, smooths are penalized for wiggliness. Crucially, the data is not averaged over participants and items so that the modeler can work with the original, complete data.

There are several advantages to using GAMMs over traditional analysis methods used in experimental linguistics, such as ANOVAs. One obvious advantage is not having to average the data, which allows the modeler to gain richer and more precise insights into the cognitive processes taking place during reading, and it is possible to model a random-effects structure (Baayen et al., 2008; Bates et al., 2015). Second, it is possible to model nonlinearities, so that the modeler is not restricted by the assumption that the relationship between the dependent variable and the predictor variables is linear. Moreover, by including predictors such as the trial number, one can control for the changes in attention of participants over the course of the experiment (Baayen et al., 2017a).

GAMMs have been applied to several linguistic datasets in previous research, including experimental studies using EEG (De Cat et al., 2015; Hendrix et al., 2017; Kryuchkova et al., 2012) or eye-tracking (Lensink et al., submitted).

We conducted an exploratory analysis in which we tested if and how age has an effect on several eye-tracking measures. We used the `mgcv` package (Wood, 2006) for fitting GAMMs to the eye-tracking data. To track the reading of lexical bundles over time, we looked at the first fixation durations, second fixation durations, and the number of fixations. For the number of fixations, we fitted a generalized linear model with a Poisson link.

We explored if age interacted with any of the frequency measures to see if frequencies have different impacts on different age groups. We furthermore explored if age interacted with any of the other predictors, to check for differences in reading patterns between older and younger adults.

2.3 Results

Prior to analysis, we automatically assigned the fixations to different regions of interest (the first, second, or third word) on the basis of their location on the screen. We furthermore removed fixations that landed too far from the written text. We log-transformed all frequencies and fixation durations to approach normality.

2.3.1 First fixation durations

First fixation durations provide an insight into the first processes that take place when participants read a trigram. As previous work showed that not only frequency information plays a role, but also the length of the trigram in characters, and the horizontal location of the fixation (Lensink et al., submitted, see **Chapter 4**), we also tested for inclusion of these measures here.

As the frequencies of the second word (B) and the first bigram (AB), as well as the frequencies of the second bigram (BC) and the trigram (ABC) correlated to a large extent ($r > 0.6$), we only included the one predictor of the correlated pairs that explained most of the variance in the data. For this dataset, the frequencies of the first bigram were better predictors than the frequencies of the second word, and the frequencies of the trigram outperformed the frequencies of the second bigram, so we included only the AB and ABC frequencies.

The model of the first fixation durations contains a significant effect of age group, with younger participants spending more time on their first fixations than older participants. In contrast, previous studies report that older adults' fixations are longer than those of younger adults (Rayner et al., 2006, 2009). However, the effect of age that we found here is small.

The model contains significant smooths of the current fixation position, the length of the trigram, the frequency of the last word (logFreqC), the frequency of the first bigram (logFreqAB), and the frequency of the whole trigram (logFreqABC). There are furthermore significant random intercepts per subject and experimental item (trigram), and random slopes per subject for the fixation location, the length of the trigram, and the frequencies of the first bigram. These random slopes show that participants differ significantly from each other on these dimensions. See Table 2.2 for details.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	4.5155	0.4458	10.1286	< 0.0001
difference young-old	1.0411	0.5103	2.0403	0.0413
B. smooth terms	edf	Ref.df	F-value	p-value
s(fixation location)	6.2285	7.3093	25.6189	< 0.0001
s(length)	1.0002	1.0003	130.4974	< 0.0001
s(logFreqC)	1.0744	1.1195	26.6106	< 0.0001
s(logFreqAB)	1.0003	1.0005	4.5563	0.0328
s(logFreqABC)	2.7905	3.2837	5.1029	0.0013
s(trigram)	105.0741	296.0000	0.6267	< 0.0001
s(subject)	43.8588	50.0000	636795.2705	< 0.0001
s(fixation location,subject)	45.8459	51.0000	599751.5208	< 0.0001
s(length,subject)	33.1212	51.0000	32716.1236	0.0008
s(logFreqAB,subject)	19.9314	51.0000	14907.2289	0.0003

Table 2.2: Results of the model of the First Fixation Durations.

The significant smooths are plotted in Figure 2.1. The top panels show the effects of the fixation location and the length of the trigram. The first panel shows that when the first fixation lands near the beginning of the trigram, participants tend to re-fixate quickly, likely because they cannot extract much information from the text. However, from 425 pixels to the left-hand side of the screen onwards, participants are able to receive more information from

the signal, and fixate longer[3]. The effect of the length of the trigram shows a negative slope. When a trigram is longer, there is more to be read, and participants likely want to gain more information as quickly as possible, and therefore are quick to re-fixate.

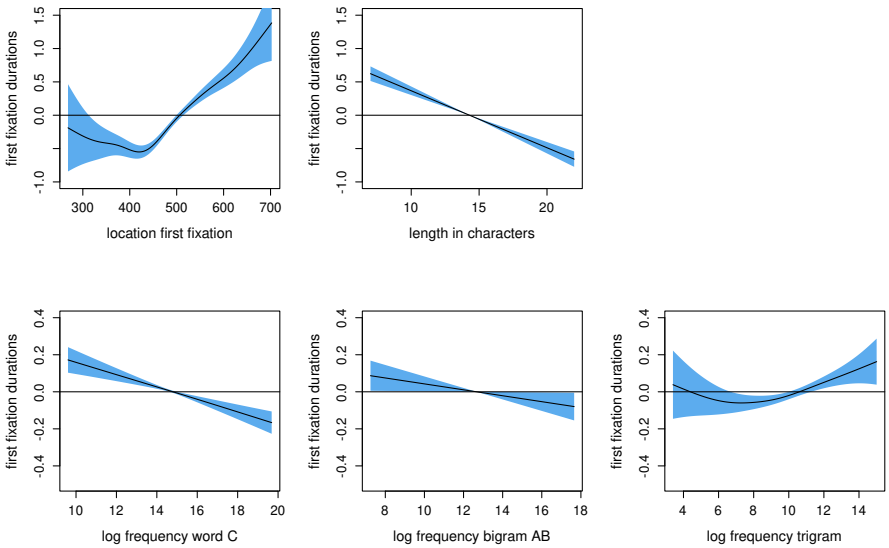

Figure 2.1: Partial effects of the model of the First Fixation Durations. The top panels show the effects of the horizontal location of the fixation in pixels and the length in characters of the trigram. The bottom panels show the effects of the frequencies of the last word, the first bigram, and the trigram.

The bottom panels of Figure 2.1 show the effects of the frequencies of the last word, the first bigram, and the whole trigram. The frequencies of the third word and the first bigram show negative slopes. That higher frequencies lead to shorter fixation durations has been reported before (Carrol and Conklin, 2015; Rayner, 1998; Siyanova-Chanturia, 2013). It is striking, however, that frequency information from the third word of a trigram already plays a role at the first fixation on that trigram, and that its effect appears to be stronger than that of the frequency of the first bigram.

The last panel displays the effect of the trigram frequencies. There is a significant effect of the trigram frequency, which shows a positive-going, slightly curved, slope. Already at the first fixation, trigram frequencies modulate reading behavior. Strikingly, higher trigram frequencies lead to longer fixations. This 'Inverted Frequency Effect' is highly unexpected. We propose that this

[3]Recall that all trigrams were left-outlined to the same location on the screen.

is a reflection of a reading strategy, where readers spend more time on the first fixation when reading something highly frequent and familiar, as they will be able to already process a large part of this high-frequency lexical bundle, without having to re-fixate. We will get back to this point in our discussion of the second fixation durations and the number of fixations, and in the general discussion of this paper.

There are no significant interactions of the different frequency measures with age. A small and probably not robust main effect of age suggests that older readers on average spend less time on their first fixations than younger readers. However, this difference is not tied to how the two groups process trigrams. Older and younger adults seem to process trigrams and their constituent parts in the same way at this stage.

2.3.2 Second fixation durations

After the initial stage of processing, readers estimate where to move their eyes next, and jump to their next fixation. In this stage, the reader has already processed some information of the trigram presented, and continues integrating previous bottom-up information, top-down expectations, and current input. All these processes are reflected in the second fixation durations.

Out of the total of 13,960 second fixations made, 6,069 (43.5%) constitute regressions. To take into account that the second fixation could be a regression or a forward fixation, we included a factor specifying whether or not the fixation is a regression (called 'regression' in Table 2.3).

The model contains a main effect of regression, and two interaction terms with the locations of the first and second fixations and the regression factor. The model furthermore includes the length of the trigram, the frequencies of the last word and the trigram, and trial number. All predictors except for trial number are significant. We also included random intercepts for items (trigrams) and subjects, and random slopes of the fixation positions, the length of the trigram, the frequencies of the last word and the trigram, and random smooths for trial per participant. The significant random slopes indicate that there is a significant amount of individual variation in how these measures influence processing in different participants.

All main smooth terms are displayed in Figure 2.2. The first two panels show the interaction of regression with the locations of the first and second fixations. The interaction effect is clearest in the first panel. If the second fixation is a regression (red color), then this second fixation lasts longer the further the first fixation landed into the trigram. Because the perceptual span is asymmetrical with a larger viewing range to the right of the fovea (Rayner, 1998), participants can benefit less from fixations near the end of the trigram. Therefore, if the first fixation landed far into the trigram, where less information is available, then participants benefit more from a regression, where they will then spend more time.

If, however, this second fixation is a forward fixation (blue color), then the

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.4774	0.1482	36.9602	< 0.0001
regressionY	0.0971	0.0894	1.0861	0.2775
B. smooth terms	edf	Ref.df	F-value	p-value
s(firstFixX):regressionN	4.6652	5.7172	4.2490	0.0004
s(firstFixX):regressionY	2.4079	3.0850	26.2545	< 0.0001
s(secondFixX):regressionN	2.1114	2.7240	8.4785	0.0001
s(secondFixX):regressionY	6.0018	6.6730	8.6134	< 0.0001
s(length)	2.8702	3.4394	32.1737	< 0.0001
s(logFreqC)	2.9538	3.5100	9.6736	< 0.0001
s(logFreqABC)	1.0004	1.0007	15.4801	0.0001
s(trial number)	1.0003	1.0005	2.0256	0.1547
s(trigram)	70.0836	297.0000	0.3222	0.0003
s(subject)	36.3351	51.0000	68576.2413	0.0002
s(firstFixX,subject)	23.7373	51.0000	18384.3967	0.0550
s(secondFixX,subject)	44.2694	51.0000	84933.8703	< 0.0001
s(length,subject)	29.3250	51.0000	13099.5019	0.0002
s(logFreqC,subject)	18.1087	51.0000	4495.9120	0.0483
s(logFreqABC,subject)	17.5828	51.0000	3799.1280	0.0090
s(trial number,subject)	31.3173	51.0000	2029.1476	< 0.0001

Table 2.3: Results of the model of the Second Fixation Durations.

effect almost flips: The further the first fixation landed into the trigram, the less information participants can gain from a forward fixation, and they will quickly re-fixate. Note that this does not hold if the first fixation landed near the beginning of the trigram — in that case, a second fixation further into the trigram does provide new information, leading to longer second fixations.

The interaction of regression and the location of the second fixation is less clear. Overall, when the second fixation is not a regression (blue color), fixations further into the trigram will last longer. If the participants did regress (red color), then the overall trend seems to be that the second fixation will last a bit shorter the further this fixation lands into the trigram. If a regression does not land far back, but more towards the end of the trigram, then not much new information can be gained, and a participant will not feel the need to spend more time than necessary at that fixation.

The third panel from the top row displays the effect of the length of the trigram in characters. Like in the case of the first fixation durations, longer trigrams take longer to read.

The bottom two panels of Figure 2.2 show the effects of the frequencies of the last word and the trigram itself. The frequency of the last word is small, and facilitative: The durations of the second fixations are shorter when the last word of the trigram is more frequent. As for the trigram frequencies, we are again seeing an Inverted Frequency Effect as in the first fixation durations:

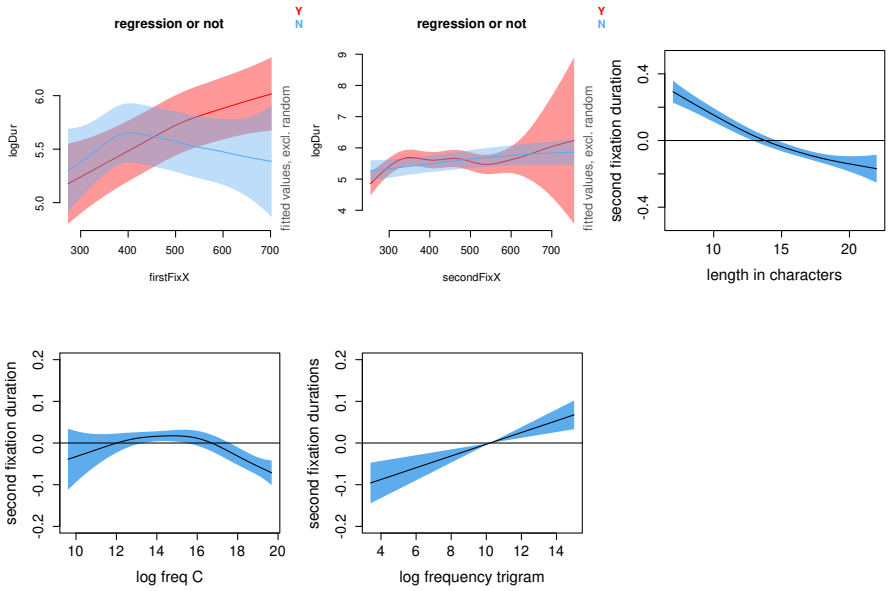

Figure 2.2: Partial effects of the model of the Second Fixation Durations. The first two panels in the top row show the interactions of the first and second fixation locations with a factor specifying whether or not the second fixation was a regression. The third top panel shows the effect of the length in characters. The bottom two panels display the effect of the frequencies of the last word and the trigram.

The more frequent a trigram is, the longer the fixation will take. As it is a widely-reported result that higher frequencies correlate with shorter processing latencies (see e.g. Rayner, 1998), this is highly unexpected.

Discussion early reading measures

Both the models from the first fixation durations and the second fixation durations show surprising and unexpected results in the form of an Inverted Frequency Effect. It is well-known and widely accepted that processing becomes increasingly easier the higher an item's frequency is (Harley, 2013). Moreover, shorter looking times are thought to reflect easier processing (Rayner, 1998; Reichle et al., 1998, 1999), so high frequency items are expected to correlate with short looking times. However, we have found exactly the opposite in our data. Higher trigram frequencies correlate with longer fixation durations. How to reconcile our findings with decades of psycholinguistic findings?

Besides the duration of fixations, the number of fixations made is also an indicator of ease of processing. The easier it is to process a (string of) word(s), the fewer fixations are needed to fully process it (Rayner, 1998). When people need few fixations to read and process a multi-word unit, because of its high phrasal frequency, they might strategically spend more time at their early fixations. Due to the high frequency of the item, readers will recognize the item early in processing, and will spend more time at this early stage of processing because they aim to process as much information as possible as early as possible.

For a low-frequency multi-word unit, however, more fixations will be needed, and people are expected to spend less time on their first fixations. As the item is not as frequent, most readers will not recognize the item straightaway, and therefore readers will tend to re-fixate quickly to gain more information from next fixations. This hypothesis can be tested by considering the role that the early fixation durations play in the number of fixations made. If it is indeed true that easier processing leads to longer first and second fixations but fewer fixations in total, then there should be a negative relationship between the length of the first and second fixations and total number of fixations made. We will explore this possibility in the next section.

2.3.3 Number of fixations

The number of fixations participants make, reflects the overall difficulty of processing (Rayner, 1998). The more fixations a participant makes, the harder it was for him to process the written text. This measure provides a summary of overall processing, as it aggregates the whole time course of reading of the presented stimulus. It can test our hypothesis that longer first and second fixations actually indicate ease of processing: If longer first fixations correlate with fewer fixations in total, then overall the whole trigram was easy to process,

and participants actually already used the first fixations to take in and process as much information as possible.

The model did not show any significant main effects of or interactions with age. This means that younger and older adults do not differ in how many fixations they make on frequent lexical bundles. There are significant smooths of the durations of the first and second fixations, the frequency of the first word, and a near-significant effect of trial number. There are moreover significant intercepts of items (trigrams), and significant random slopes of the frequency of the first word per subject, indicating that the reaction to the first word frequency varied a lot among the participants.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1.3691	0.0146	94.0710	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(duration first fixation)	6.7437	7.6910	759.3490	< 0.0001
s(duration second fixation)	4.6009	5.6334	821.3799	< 0.0001
s(logFreqA)	1.0000	1.0000	7.3502	0.0067
s(trial number)	1.0000	1.0001	2.8522	0.0913
s(trigram)	266.4023	299.0000	2785.6496	< 0.0001
s(subject)	0.0025	51.0000	0.0025	0.4890
s(subject, dur first fix)	0.0002	51.0000	0.0001	0.8585
s(subject, dur second fix)	0.0003	51.0000	0.0003	0.6860
s(subject, logFreqA)	32.2192	51.0000	92.2494	< 0.0001
s(subject, trial number)	0.0001	51.0000	0.0001	0.9403

Table 2.4: Results of the model of the Number of Fixations.

The plots in Figure 2.3 show how longer durations in the first and second fixations correlate with fewer fixations overall, as we hypothesized in section 2.3.2. As fewer fixations indicate easier processing, it must be the case that readers prefer to spend more time at their early fixations when an item is easy to process, while making fewer fixations overall. In other words, it is not necessarily the early measures that reflect ease of processing of phrasal units, but late measures such as number of fixations. This is especially relevant in light of the unexpected findings for the durations of the first and second fixations, where higher trigram frequencies result in longer looking times.

It is surprising to see that the first word frequency is the only frequency information that significantly influences the number of fixations — especially since this first word frequency did not play a role in the earlier measures. Most of the first words of the trigrams used in this study are function words (93%). It seems unlikely that high frequency function words would aid in processing, as high frequency function words typically can be followed by a much larger set of words than low frequency function words. This would make processing harder, instead of easier. Because of the unbalanced distribution of function words and content words at the first position of the trigram in our set of stimuli, it is

important to run a replication study where this distribution is more balanced.

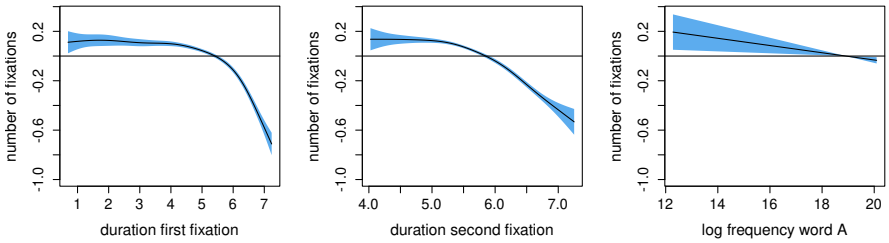

Figure 2.3: Partial effects of the model of the Number of Fixations. The first two plots display the effects of the durations of the first and second fixations. The third plot show the effect of the frequency of the first word of the trigrams.

2.4 General discussion

To study how the processing of written lexical bundles proceeds over time, we tracked the eye-movements of older and younger people reading a large set of high-frequency Dutch trigrams. We expected to see an age-related difference in the processing of multi-word units. Specifically, we expected that the influence of the trigram frequencies would differ between the age groups, but we were uncertain how that difference would manifest itself. We predicted to either see an increased reliance on multi-word units in older adults, or a decreased reliance on multi-word units and an increased reliance on combinatorial processes. However, we found no age-related differences in how trigrams are processed, but only quantitative differences between the age groups in the first fixations, where older adults tend to spend less time than younger adults. This is unexpected as older readers are known to read slower (Laubrock et al., 2006; Rayner et al., 2006, 2009). However, the effect is small and probably not robust, as we did not find any effects of age in the second fixations and the number of fixations made.

There is a possibility that the results do reflect an increased reliance on lexical bundles in older adults. As older adults are known to have a larger lexicon (Keuleers et al., 2015), and since searching through a larger lexicon is more time-consuming, resulting in slower latencies (Ramscar et al., 2014), not finding any slower latencies in lexical bunding reading might actually indicate a facilitative effect of lexical bundles in older adults. Still, it is not possible to establish any findings from a null-effect. Future studies, both experimental and computational, could perhaps shed more light on the validity of this speculation.

The non-linear mixed-effects modeling approach has laid bare a precise overview of how people read lexical bundles over time, with the surprising findings that readers seem to employ a strategy of spending more time on their early fixations when reading a highly frequent lexical bundle. The models of the first and second fixation durations and the number of fixations show an intricate interplay of several predictors, among which the locations of the fixations themselves, the location of the previous fixation, a factor specifying whether or not a fixation is a regression, the length of the trigram in number of characters, and several frequency predictors. Moreover, the inclusion of random effects structures has shown that there are substantial individual differences in how readers process the trigrams.

From the first fixations onwards, several frequency measures play a role in processing, including the trigram frequency. From the first fixation onwards, readers already manage to identify which lexical bundle they are reading. This is similar to findings in Miwa et al. (2017), who found whole-form frequency effects of trimorphemic compounds already in the first fixations, and to Lensink et al. (submitted), where full-form trigrams played a role from the first fixation onwards.

At first, the trigram, the first bigram, and the last word play a role in processing. At the second fixation, the effect of the first bigram frequency has worn off, and the effect of the frequency of the last word has been attenuated. It seems to be the case that readers try to take into account several levels of information at once at the beginning — the single words, the bigrams, and the trigram — whereas later onwards, they focus more on further confirming what they have already processed, and are more strategic in where they move their next fixations. The intricate pattern of interactions of fixation locations and regressions testifies to the latter.

2.4.1 The Inverted Frequency Effect

The shape of the effect of the trigram frequencies is puzzling — higher trigram frequencies lead to longer fixations, whereas higher unigram and bigram frequencies lead to shorter fixations. Lensink et al. (submitted) also found an inhibitive effect of trigrams in a reading task. As shorter fixations indicate ease of processing (Rayner, 1998), and as higher frequencies items are easier to process, this result is unexpected. Moreover, modeling studies of reading behavior (Reichle et al., 1998, 2012) have shown repeatedly that higher frequency items correlate with shorter reading times, also in older readers (McGowan and Reichle, 2018).

In order to check if we are not dealing with an effect of suppression or enhancement (Wurm and Fisicaro, 2014) that caused the sign of the effect to flip, we inspected the results of a model with trigram frequency as the sole predictor, and found that the trigram frequencies still have an inhibitive effect on fixation durations. We also checked the correlation between trigram frequency and the first fixation duration, and found a positive correlation, r

= 0.05, again showing that the direction of the effect is probably not due to suppression or enhancement.

We have referred to this inhibitive effect of trigram frequencies as the Inverted Frequency Effect. It is reminiscent of the Inverted Optimal Viewing Position Effect (Vitu et al., 2001), where fixations at the center of words — thought to be the optimal place for processing, and thus expected to correlate with shorter fixations — elicit longer fixations. An explanation for this effect could be that Letters in the center of words have more interference from neighboring letters than letters at the margins.

As for the Inverted Frequency Effect, it is possible that the effect of trigram frequencies does not reflect ease of processing, but competition between similar trigrams. Higher frequency trigrams are part of a larger group of similar multiword units, and those might have an inhibitive effect, resulting in longer looking times. A competition account is consistent with the timing of the effect, early in processing, where lexical access and thus competition could still take place. However, this does not explain why only trigrams would be affected in this way — neighborhood density effects have been reported for single words too (Baayen, 2010; Yates et al., 2008).

An alternative explanation is that high frequency trigrams activate a larger semantic network of associated meanings, senses, and synonyms, which might slow down processing. Again, this does not explain why single words and bigrams are affected differently. It could be that enlarged semantic processing is only elicited by phrasal elements. The trigrams used in this study are mostly constituents *aan het begin* 'at the beginning'. If it is indeed the case that more phrasal units elicit another type of processing than non-phrasal combinations of words, then it is expected that there will be a clear difference in processing between phrasal lexical bundles and non-phrasal lexical bundles — a prediction that has yet to be tested.

A third possibility, i.e. a reading strategy, was tested in Section 2.3.3. We hypothesized that readers prefer to spend more time at their early fixations when an item is easy to process, while making fewer fixations overall. Indeed, we found that fewer fixations correlate with longer first and second fixation durations. The more time readers spend at their early fixations, the fewer fixations they will need overall. Therefore, it is not necessarily the early measures that reflect ease of processing, but late measures such as number of fixations. Again, like in the competition account and the enlarged semantic processing account, it is not clear why only trigrams show an Inverted Frequency Effect.

2.4.2 Limitations and future directions

In this study, we looked at two groups: Younger and older readers. However, by dichotomizing the age variable, it is not possible to know if these differences change gradually over time, or if there are one or more abrupt changes. Also, the age split is chosen arbitrarily, and might not align with what the real distribution looks like in the population, which in turn could increase the probability

of finding spurious results (McWhinney et al., 2016). Moreover, as previous research has shown that the morphological complexity of a language has a bearing on whether or not age-related differences can be found (Reifegerste et al., 2017), it could be that age-related differences in lexical bundle processing can be found in other languages than Dutch.

Furthermore, it might be the case that the age gap between the two groups was not large enough to find age-related differences in the processing of lexical bundles. For children, whose lexicon is a representation of a relatively small set of experiences, each new instance of a word or lexical bundle will have a relatively large impact on the form of the representation of that word or lexical bundle. However, for an adult, the impact of each new instance of a word or lexical bundle will be much smaller, as their present representations are an accumulations of a much larger set of experiences and as such are much more stable and robust than the representations a child has. Therefore, the differences in representations of lexical bundles are expected to be much larger between children and adults than between younger and older adults.

Moreover, if we assume that representations of lexical bundles in memory are proportional to the log frequencies of these lexical bundles, then using only high-frequency stimuli might obscure differences between younger and older adults[4]. Suppose that a certain high-frequency item has occurred 20,000 times in the lifetime of a twenty-year-old, and 60,000 times in the life of a sixty-year-old. The ratio of the log of these frequencies, ln(60,000)/ln(20,000), is roughly 1.1. However, the ratio of a low-frequency item that has occurred only 200 times for a twenty-year-old and 600 times for a sixty-year old, is roughly 1.2. This larger difference in log occurrences of low-frequency items might make it more likely to detect a difference in age groups, assuming that even low-frequency lexical bundles play a role in processing (and there is evidence suggesting they do (Arnon and Snider, 2010)). Still, the difference in ratios is not that large, and phrasal frequency effects are much larger for high-frequency items.

The set of stimuli used in this study is limited in its distribution of word categories and trigram frequencies. A large majority of the first words of the trigrams are function words, which could skew the results. Moreover, all lexical bundles have been sampled from the top one per cent most frequent lexical bundles found in the OpenSoNaR corpus. Still, the present study provides a starting point for future studies investigating age-related differences in lexical bundle processing.

Therefore, for future studies, a bigger group consisting of participants of a more diverse and wider range of ages should be tested, specifically including children. One can then test if the absence of any age-related effects is robust, and if not, it will be possible to study how the processing of lexical bundles changes over the years, if the change over time happens gradually or abrupt, and, if the latter applies, from what age onwards these changes take place. Future studies should furthermore take into account the distribution of function

[4]We thank an anonymous reviewer for this suggestion

and content words, if the lexical bundles are phrasal units or not, and should sample from a (much) wider frequency range. These studies should also control for the neighborhood density of the lexical bundles used, as this might also cause competition and thus inhibitive effects.

2.5 Conclusion

For the present set of stimuli, we did not find any age-related differences in how younger and older adult readers process frequent Dutch lexical bundles. Still, the use of non-linear mixed-effect regression models has allowed us to study how several linguistic and oculo-motor features play a role and how they interact when people are reading frequent lexical bundles. Trigram frequencies are an important factor at the early stages of reading. Interestingly, trigram frequencies show an Inverted Frequency Effect, where higher frequency trigrams correlate with longer looking times. These longer looking times for early fixations in turn correlate with fewer fixations overall — an indication of ease of processing — which suggests that readers strategically spend more time at the first fixations when an item is easy to process. These findings show that it is not necessarily the early measures that reflect ease of processing of phrasal units.

Acknowledgement(s)

The authors would like to thank Sabine Nauta and Amber de Bruijn for their assistance in collecting the data, and an anonymous reviewer for comments and suggestions that have improved the manuscript.

CHAPTER 3

Listening

Processing spoken multi-word units: an ERP investigation

Saskia E. Lensink, Antoine Tremblay, Lilian Ye, Arie Verhagen, Niels O. Schiller

abstract
We studied the on-line processing of auditorily presented lexical bundles. Participants were presented with a set of high-frequency lexical bundles and matched controls, while EEG data was collected. We found a sustained early negativity with an early onset that was more pronounced for the matched control items. The data were analyzed using conditional inference random forest modeling (CForest) to gain detailed insights into the time course of auditory processing of lexical bundles, the possible neural sources recruited over time and linguistic and non-linguistic factors that mediate auditory processing. We propose there are three stages that are reminiscent of single word comprehension, representing 1) predictive and bottom-up processes; 2) inhibition and competition; 3) lexical integration. The data provide evidence for an interactive processing model.

Keywords ERPs, multi-word units, auditory processing, comprehension, conditional forest modeling

3.1 Introduction

There is a growing body of work suggesting that language users are sensitive to phrasal frequencies (Bannard and Matthews, 2008; Shaoul and Westbury, 2011; Siyanova-Chanturia, 2013). This is not surprising when considering idioms such as *kick the bucket*, where the meaning of the single words combined does not equal the meaning of the whole. However, phrasal frequency effects are also found for frequent, completely regular, and transparent combinations of words (Arnon and Snider, 2010; Tremblay and Tucker, 2011). These combinations are often referred to as 'lexical bundles'. Examples of lexical bundles are *on the day* or *I think that*. These combinations are thought to play a role in processing because of their common co-occurrence, which encourages the brain to chunk these words together into building blocks of language (Bybee, 2006; Green, 2017).

There is a lot of experimental evidence that phrasal frequencies play a role in processing when speaking and reading. However, it is not yet clear if these phrasal frequencies are also important when listening to language. Moreover, little is known about the time-course of processing of common combinations of words. This study aims to fill this gap by presenting an EEG study on the on-line processing of auditorily presented lexical bundles.

3.1.1 Previous work on the time course of multi-word unit processing

Previous studies have sought to understand the time-course of lexical bundle processing by running eye-tracking and EEG experiments. The electrophysiological signal of the brain can be used to derive Event Related Potentials or ERPs, which are reflections of on-line processing unfolding over time (Kutas and Van Petten, 1994). Only a small number of studies has used ERPs to investigate frequent combinations of words, and most of those studies focused on idioms. In a reading study, Vespignani et al. (2010) compared ERPs elicited by idioms and literal sentences. Past the recognition point of the idiomatic phrases – the word past which participants could recognize the phrase as being idiomatic – idioms elicited an enlarged P300 as compared to their matched literal phrases. The P300 has been found when participants are presented with highly predictable items, such as the lexical item *white* after the presentation of the sentence *The opposite of black is ...*, (Roehm et al., 2007), or the correct answer to a simple calculation (Fisher et al., 2010). The presence of a P300 in the Vespignani et al. (2010) study shows that after the recognition of an idiomatic phrase, items completing the idiom are actively predicted and pre-activated.

Two previous studies have looked at ERPs of lexical bundles (Hendrix et al., 2017; Tremblay and Baayen, 2010). Tremblay and Baayen looked at the electrophysiological signal of participants reading regular four-word sequences. The whole-string frequencies of these sequences ranged from anywhere between very low (0.01 per million) to very high (100 per million). The authors found that a higher whole-string probability corresponded to a more negative N1 and a less positive P1. The N1 and P1 are early ERP components occurring just 100 ms after stimulus presentation. As the earliest reported frequency effects of single words occur around 100 ms after stimulus onset (Hauk et al., 2006; Penolazzi et al., 2007; Sereno et al., 1998), Tremblay and Baayen reasoned that it would not be possible for multiple words to be accessed and combined within this short time frame. Therefore, they argued that their results show that four-word sequences are retrieved holistically, as if they were a single word.

A couple of years later, Hendrix et al. (2017) investigated the online processing of lexical bundles by presenting participants with a prime consisting of a preposition plus a definite article, followed by a picture of a concrete object. The prepositional phrases had different phrasal frequencies. The authors found effects of both single word frequencies and phrasal frequencies during the naming of the object. Effects of single word frequencies were already present 95 ms after stimulus presentation and occurred mostly in the left hemisphere. Effects of phrasal frequencies were seen as a sustained negativity over the left hemisphere, with higher frequencies correlating with more negative voltages. Hendrix et al. (2017) argue that the different ERP patterns observed are evidence that words and phrases are processed differently. Note that Tremblay and Baayen (2010) did not find any sustained negativities in their study, but only

found more negative voltages in early ERP components for higher frequency lexical bundles. Note, however, that Tremblay and Baayen (2010) looked at reading, whereas Hendrix et al. (2017) focused on speaking.

Another way to study the time-course of on-line processing is by employing eye-tracking methods. Eye-tracking provides an indirect means of investigating in which order parts of words and sentences are processed and which cognitive processes might be involved, the idea being that eyes focus on the item that is being processed, and that the duration of gazes indicates the ease of processing (Just and Carpenter, 1980). Recently, Lensink et al. (submitted) used eye-tracking to study the on-line processing of lexical bundles. In line with previous research, Lensink et al. found that more frequent lexical bundles are easier to process than less frequent ones. Moreover, the authors found evidence that lexical bundle frequencies already play a role in gaze durations of the first fixation, showing the early onset of phrasal frequency effects in reading.

These previous studies had participants reading silently or aloud from a screen. There are several disadvantages to studying reading behavior in this way. People can have different reading strategies, and the researcher has little control over the order of processing of the item presented. There are also pitfalls to studying the production of lexical bundles, where participants either have to first read from a screen, or recall items from a list, before they start to articulate. Some participants start talking as soon as they have identified or recalled the first word, whereas others might wait until they have read or recalled the whole string. These different strategies are likely to originate from different cognitive processes, which in turn have different effects on the way participants produce speech. However, there is an alternative where the researcher can precisely control and track the order in which participants receive the input: listening.

Many ERP studies investigating the auditory processing of speech study the time course of phonological, semantic, and syntactic processing, and the influence of context. This is done by presenting participants with full sentences that are either correct, semantically anomalous, or syntactically anomalous. Semantic errors are mostly reflected in a larger N400, and syntactic errors are reflected in larger left anterior negativities (LAN) and a more positive P600 (Friederici, 2002). Oftentimes an enlarged early negativity is also found. Some researchers interpret this negativity as a marker of a phonological mismatch between what is expected and what is heard (phonological mismatch negativity (PMN); Connolly and Phillips, 1994), whereas others consider it a marker of initial form-based assessment of the incoming signal (N200/250; Hagoort and Brown, 2000; Van Den Brink et al., 2001; Van Den Brink and Hagoort, 2004), or a marker of word category violations (ELAN Friederici, 2002; Steinhauer and Drury, 2012). Besides this early negativity, sometimes a sustained negativity with an early onset is seen in auditory processing (Holcomb and Neville, 1991). Although some of these sustained negativities might originate from spill-over effects due to the processing of the previous word (Mueller et al., 2005; Steinhauer and Drury, 2012), they have been linked to working memory processes in several studies (see e.g. Steinhauer et al., 2010).

3.1.2 Current study

If lexical bundles are used in processing, and if listeners make use of them during listening, then it is expected that there is a difference between the ERPs elicited by lexical bundles and ERPs elicited by matched control phrases. Importantly, this difference is expected to arise from the moment that listeners notice that they are listening to a lexical bundle, instead of just any combination of frequent words.

As soon as a listener strongly suspects she is listening to the first part of a lexical bundle, she will expect to also hear the last word of that lexical bundle. When indeed the utterance continues as expected, this match between the expected and the observed might elicit a P300. However, if a listener hears a different word than expected, his expectations are violated. This, we predict, could lead to a larger N400 component at the unexpected word. The N400 is sensitive to frequency and predictability information and has a more negative amplitude when the frequency is lower or the item is less predictable.

Another possibility is that we will see a slow anterior negativity for infrequent combinations.The amplitude of the slow anterior negativity is thought to reflect the amount of resources devoted to short-term memory processes (Kluender and Kutas, 1993; Steinhauer et al., 2010). Retrieving and combining multiple items from the mental lexicon is likely to require more working memory than retrieving a single item or a lexical bundle directly from memory. This could be be reflected in less negative sustained anterior negativities.

In what follows, we will present our exploratory analyses on the time course of auditory processing of lexical bundles. We will first discuss our methods and materials in Section 3.2. In Sections 3.3 and 3.4, we will present the results and will discuss their implications in Section 3.5.

3.2 Materials and Methods

3.2.1 Participants

We recruited forty Dutch native speakers (ten males, mean age 21.4 years) for this study. All participants had normal or corrected-to-normal vision and none of them reported any hearing deficits. After the experiment, participants received a small financial reward for their time.

Two participants were excluded due to technical issues during the experiment, and two participants were excluded because their score on the Edinburgh Handedness Inventory questionnaire (Oldfield, 1971) was negative, indicating left-handedness. The remaining 36 participants (ten male) were on average 21.3 years of age.

3.2.2 Stimuli

We created twenty-six trigram pairs where we contrasted a high-frequency multi-word unit (MWU) with a matched control (Control). The high-frequency multi-word units consisted of three words and were randomly sampled from a set of 1,000 high-frequency trigrams found in the Dutch *Ten Ten* web corpus (Jakubíček et al., 2013).

Matched controls were made by taking a high-frequency trigram and changing its last word, thereby creating a low-frequency trigram. Crucially, we made sure that, for each MWU and Control pair, the final words had similar frequencies, but that the phrasal frequencies were different by at least a factor ten. To give an example, we used the trigram *een belangrijke rol* ('an important role') and changed its last word to create the trigram *een belangrijke vorm* ('an important form'). These trigrams differ only in their phrasal frequencies, whereas all single word frequencies are similar. This way, we can disentangle the effects caused by the terminal single words and the effects of the phrasal frequencies of the whole trigrams. Similar sets of stimuli have been used in several studies looking into multi-word units (see for example Arnon and Snider, 2010).

To ascertain that our multi-word units had a different phrasal frequency than their matched controls, we checked for prevalence of each stimulus pair in different corpora, i.e. the Dutch *Ten Ten* web corpus (Jakubíček et al., 2013), the *Europarl* corpus (Koehn, 2005) and the *EUR-lex* corpus (Baisa et al., 2016). We extracted the frequencies of the trigrams and their constituent words from the Netherlands Dutch subset of the OpenSonar corpus (Oostdijk et al., 2013). The stimuli and their frequencies can be found in Appendix A.

Besides the target items, we included another 52 trigrams that served as filler items. This resulted in a total of 104 different Dutch trigrams, which we pseudo-randomly put into two different experimental lists. We made sure that there was no semantic or phonological overlap between trigrams within at least two consecutive trials. Furthermore, we took care in inserting at least twenty trials between any MWU and Control pair, to minimize the likelihood that participants would recognize the similar form of *een belangrijke rol* and *een belangrijke dag*. The experiment was built in the experimental presentation software E-Prime 2.0 (Psychology Software Tools).

We recorded a male voice reading out loud the stimulus and filler items using a portable USB 2.0 Audio Interface Quad-Capture UA-55 (Roland) at a sampling rate of 44,000 Hz in mono. We created a list where our MWUs, Controls and fillers were randomly presented. We did not inform the speaker about the intention of our study prior to the recordings. Afterwards, we edited the recordings in Praat (Boersma and Weenink, 2016), adding a 500 ms silence before the onset of each stimulus and scaling all stimuli to an equal intensity of 70 dB. There were no significant differences between the acoustic durations of the control trigrams and the multi-word unit trigrams ($p = 0.6392$).

3.2.3 Procedure

Before the start of the experiment, participants completed a questionnaire on their (linguistic) backgrounds, and they filled in the Edinburgh Handedness Inventory test to check to what extent they were right-handed. All participants gave written informed consent before starting the experimental procedure.

Participants were seated in a quiet and sound-proof room in front of a computer screen. In the room, two audio boxes were placed at the front-left and front-right corner. Answer buttons were present on both armrests of the chair in which the participants were seated. Behind the chair, BioSemi ActiveTwo EEG recording equipment was placed. Participants were all connected to a 32 channel EEG set-up while the experimenter explained the experimental task detailed below.

The experiment started with an instruction screen, which was followed by a short practice block where participants could familiarize themselves with the task, and where the experimenter could check if all audio equipment was working properly. The experiment consisted of a practice block of four trials and two experimental blocks of each 52 trials, separated by a short break. The whole experiment took about ten to fifteen minutes to complete.

Each trial lasted three seconds. It started with a fixation cross that appeared in the middle of the screen for 250 ms. Then, after a silence of 500 ms, a trigram was presented auditorily through the audio boxes. To ensure that participants kept paying attention to the task, one third of the auditorily presented trigrams was followed by a visually presented follow-up phrase. All texts were presented in Courier New, font size 12, in black, on a white background. Participants had to judge whether the follow-up phrase could be a grammatical continuation of the trigram that they had just heard. For a correct answer, they had to press the button on their left, and for an incorrect answer, they had to press the button on their right. The words 'correct' and 'incorrect' were also printed on the left-hand side and the right-hand side of the screen to aid the participants.

3.2.4 EEG recordings

The EEG was recorded using 32 Ag/AgCl electrodes (BioSemi ActiveTwo), which were placed on the scalp sites according to the standards of the American Electroencephalographic Society (1991). We monitored eye movements with four flat electrodes, two of which were placed above and below the left eye, and the other two were placed to the sides of both eyes. Another two flat electrodes were placed behind the ears, at the mastoids, to monitor jaw movements. We used the CMS and DRL electrodes as our ground reference and sampled the EEG signal at 512 Hz. Afterwards, the EEG signal was re-referenced off-line to the mean of the two mastoids and band-pass filtered (0.05-30 Hz) in Brain Vision Analyzer (version 2.0). Eye blinks were corrected by means of an ICA procedure.

Auditory comprehension studies are known to be susceptible to spill-over

effects due to processing differences of the words prior to the target words. Late ERP components of a word, such as the N400 or P600, may cause artifacts in the ERPs of a subsequent word if the words are in close proximity in time. Also, even lexically identical words may differ prosodically and phonetically when they are followed by different words, leading to co-articulatory differences, which in turn could lead to differences in the EEG signal spilling over into the target word (Steinhauer and Drury, 2012).

Recall that the stimulus items only differ in their last words, but that the second word might contain articulatory traces of this last word. To control for these effects, and to also control for any spill-over effects due to the processing of the second word, we time-locked the ERPs to the onset of the last syllable of the second word and performed a 200 ms baseline correction. We choose for the onset of the last syllable of the second word instead of the onset of the second word, as the number of syllables of the second word differed across stimuli, with two-thirds of the stimuli having a one-syllable closed-class word in second position. We also reasoned that any co-articulatory effects would be most perceptible in the last syllable.

3.2.5 Conditional inference random forest analysis

We analyzed the EEG data using conditional inference random forests (CForests). Random forests are a widely used machine learning algorithm that can be used for both categorization and regression analyses. CForests have been gaining popularity in fields such as genetics, epidemiology, and medicine, and, more recently, have been applied to several (neuro)cognitive and psychological datasets (McWhinney et al., 2016; Strobl et al., 2009) and linguistic data (Tagliamonte and Baayen, 2012).

A random forest algorithm repeatedly splits the data into two groups based on a set of predictors. The first split is made by testing which predictor explains the most variance in a random subset of the data, and then by determining at which level or value of this predictor the subset can be split into two. The algorithm continues splitting the data until it cannot find any significant features that would warrant any further splitting. The result of this first set of steps is a hierarchical structure know as a classification tree.

The name random forest is chosen because the algorithm does not create a single classification tree, but a large set of classification trees, each based on a different random subset of the data. The final model is based on an average of the predictions of the forest. See Strobl et al. (2009) and Hothorn et al. (2006) for a more detailed discussion.

There are several advantages to using CForests over more traditional parametric methods such as mixed-effects regression. CForests are non-parametric models that do not assume that the data follows a specific distribution. They can model any type of (non)linear relations between predictor variables and outcome variables and are very robust to noise. Another significant advantage is that most of the modeling process is data-driven instead of dependent on

human decisions. The modeler does not need to define in advance what shape the functional relation between the predictors and the outcome variable has, nor does she have to define which interactions have to be tested. Any strong simple or complex higher-order interactions that are present in the data will be picked up by the model itself.

Whereas the results from both forward and backward model fitting in regression modeling are notoriously susceptible to the order in which predictors are added or deleted (Strobl et al., 2009), CForests do not suffer from this drawback as they represent an aggregated average of a diverse set of classification trees - each of which is built on a random subset of the data and can therefore take on any type of form. It is important to keep in mind that CForests are truly random models in that there might be slightly different results every time the model is run. Stability and robustness are established by growing a large set of trees. Small effects that might go undetected in parametric regression methods, could still surface in some of the trees in the forest and appear in the aggregated results.

In this study, a total of 10,000 trees was grown on random subsets of the data, and furthermore *variable preselection* was applied, where not only each tree is grown on a subset of the data, but each tree node is split using a random subset of the predictors. Variable preselection produces an even more diverse set of trees (Breiman, 2001). The random subsets consisted of 33.2% of the data for each tree, and for each tree, one variable was randomly selected at each tree node.

3.3 ERP results

In Figure 3.1, the ERPs measured at frontal, central, parietal, and occipital electrodes are plotted. At the frontal, central, and parietal electrodes a clear P100 component is visible, which is more positive for multi-word units at frontal and central electrodes at the midline and right hemisphere. The plots furthermore reveal a small N200 component which is most pronounced at frontal electrodes. Hagoort and Brown (2000) report on an N250 component elicited by semantically anomalous words in spoken sentences. Although we did not use anomalous words in the control items, the final word of the control items is less expected and seems to elicit the same type of response, perhaps less strongly, as semantically anomalous words.

Overall, there is a slow-going anterior negativity that seems largest at frontal areas, and which is more pronounced for control items. This negativity is similar to the sustained anterior negativities found in previous studies on the processing of continuous speech (Hagoort and Brown, 2000; Holcomb and Neville, 1991), which have been linked to increased working memory demands as a result of syntactic processing (Coulson et al., 1998; King and Kutas, 1995; Müller et al., 1997). Moreover, research has shown that the distribution of the negativities is wider and less lateralized for increased working memory conditions

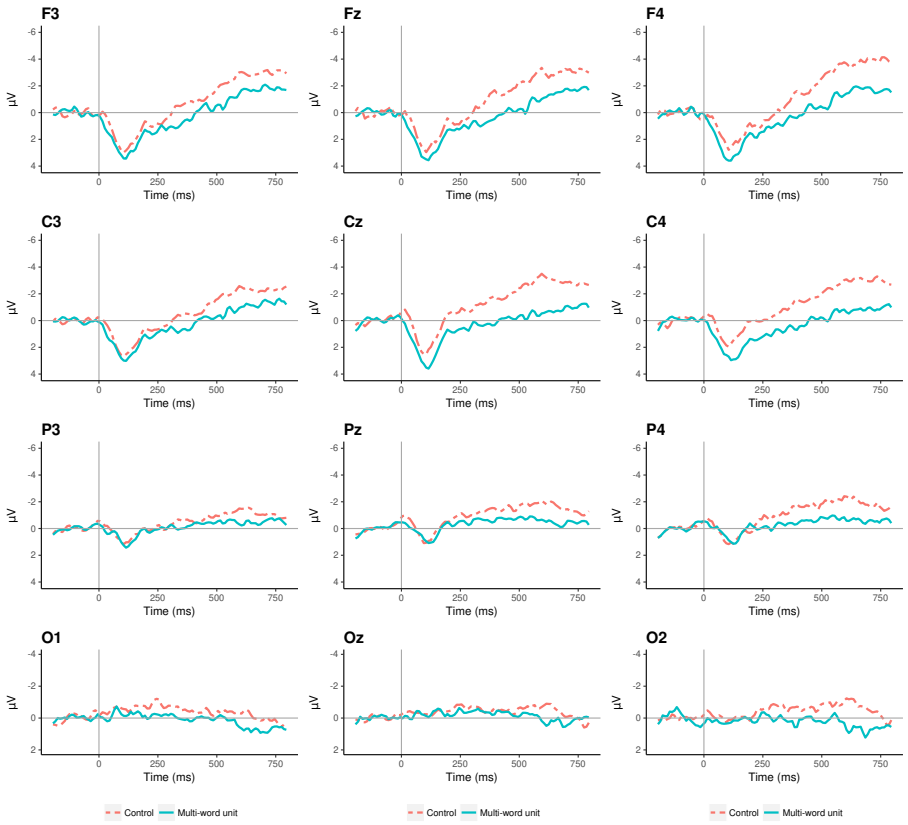

Figure 3.1: Grand-averaged ERP waveforms time-locked to the last syllable of the second word. The control condition is presented in red with a dashed line, and the multi-word unit condition is presented in blue. Presented are frontal, central, parietal and occipital electrodes at the midline, and the left and right hemisphere. By convention, negativity is plotted upwards.

than for grammatical violations (Martín-Loeches et al., 2005), suggesting that our results reflect increased working memory demands when participants are listening to control items as opposed to high-frequency multi-word units.

Multi-word units and control items start to diverge very early at frontal and central electrodes, with a divergence already visible at the P100 at the midline and right-hemisphere. At the parietal and occipital electrodes, the conditions start to diverge from approximately 200 ms after the last syllable of the second word. Overall, the divergence is widely distributed, and seems most prominent at fronto-central regions. Our results are the opposite of the pattern demonstrated by Hendrix et al. (2017), who found more negative voltages for high frequency phrases, which was moreover most prominent in parietal and occipital regions. Note, however, that Hendrix et al. (2017) used a production task. Moreover, the authors used a picture naming task, which needs involvement of the visual cortex, which could explain the more posterior distribution of their results.

3.4 CForest modeling results

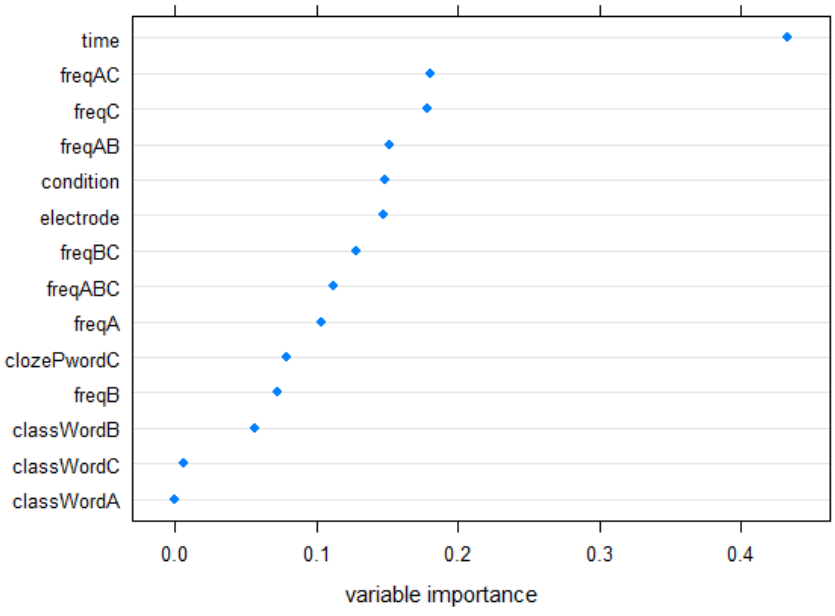

Figure 3.2: Dotplot that shows how much each predictor contributes to explaining the variance in the electrophysiological signal.

Even though interpreting an opaque model such as a conditional random forest is not straightforward, it is possible to study which variables are considered most important by the model, and to study which patterns emerge in individual trees. Figure 3.2 shows how important each variable is within the whole conditional inference forest model in describing the data. The higher on the list, the more variance in the data is explained by that variable.

The variable time is by far the most important predictor, which determines to the largest extent what voltage is generated by the pyramid cells. As the signal fluctuates quite a lot over time, this result is not surprising. A bit lower on the list is the electrode, showing that the location on the scalp also determines to a moderate extent how the signal is manifested. This is not surprising either as EEG data typically shows a lot of variation between different parts of the scalp.

The next items on the list are more interesting: Clearly, the skigram and last bigram frequencies (freqAC and freqAB), the last word frequency (freqC), and the condition (multi-word unit or control item) play the largest roles in explaining the shape of the signal. It is interesting to see that the frequencies of constituent two-word combinations have a larger impact than the frequencies of the three-word combinations themselves.

The the last bigram, the full trigram, and the first and second word frequencies (freqBC, freqABC, freqA, and freqB), the cloze probability of the last word and the class of the second word play a moderate role, whereas the word classes of the first and third words are quite small.

Although these variable importance plots can be insightful, they cannot tell us anything on the direction of an effect — i.e. does a higher bigram frequency correlate to a more negative or a more positive signal? —, nor does it show which interactions might exist — i.e. from which moment in time does the trigram frequency start to play a role? And are any frequency effects located in a specific region?

Because we are interested in learning how lexical bundle processing proceeds over time, it is worthwhile to dive into the structure of a conditional inference tree, to get a better grasp of how the different predictors interact over time. This is a trade-off: When only looking at the complete model, one has to accept that it is quite opaque and cannot provide us with in-depth and detailed insights. It will however be a quite accurate model for making predictions. However, by also looking at only a part of the model, one can conduct a detailed and in-depth study of what factors play a role, how they interact, and how they develop over time. In this article, we opt for the latter option, explicitly so to generate new hypotheses on how spoken lexical bundle processing proceeds, to guide future research.

3.4.1 Results Conditional Inference Tree

In Figures 3.3, 3.4, 3.5 and 3.6, a graphical representation of a conditional inference tree is presented. The tree-like representation shows a large number of

binary splits that produce several subgroups of data. Note that all intermediate and terminal nodes have been numbered for easy reference in the text.

The terminal nodes at the bottom of the figure represent the different sets the data has been grouped into by the model, with the number of data points and the average amplitude in microvolts of that specific group. For example, the leftmost terminal node, number seven, consists of 1970 data points (n = 1970), and its average voltage is -0.203 microvolt (y = -0.203). The binary splits are displayed in order of importance: The highest split, node number one, splits the data into two different time bins, and constitutes the strongest predictor of amplitude values. It is interesting to see that the most important binary split happens at 402 ms. The N400 component is widely reported to take place around this time, and has been connected to processes of lexico-semantic integration (Kutas and Van Petten, 1994; Steinhauer et al., 2008).

The subset of data generated before 402 ms can be found at the left-hand side of the figure. This subset, in turn, has been split into a group of fronto-centro-temporal electrodes and more parieto-occipital electrodes by node number 2. Node number 3 splits the subset of fronto-centro-temporal electrodes in a subset of data generated before 270 ms and a subset of data generated between 270 and 402 ms after the onset of the last syllable of the second word. The further one looks down the tree, the more interactions become apparent, and the way in which different factors play a role in different subsets. Therefore, the CForest analysis allows for an in-depth investigation of which factors play a role at different stages of lexical access.

In what follows, we will discuss the results chronologically and topographically, focusing on which factors play a role at different time windows and at different locations. We will relate the modeling results to what is known about the timing of lexical processing and the possible neural sources of subprocesses of lexical access. Although the source of an ERP amplitude is hard to establish on the basis of the mere location of the signal (Steinhauer et al., 2008), it is useful to speculate on its possible neural sources and what the cognitive functions of these possible sources can tell us about how processing proceeds.

As the model clearly subdivides the data into three periods, we propose that these subsets reflect three stages of lexical access. Stage 1 takes place up until 270 ms after hearing the onset of the last syllable of the second word of the trigram. Here, participants predict what might be coming next, while at the same time making use of bottom-up information. At stage 2, taking place between 270 ms and 402 ms, processes of inhibition and competition start to play a role. Then finally, at stage 3, which takes place after 402 ms, lexical integration of both bottom-up and top-down information takes place.

3.4.2 Stage 1: Prediction and bottom-up information, 0-270 ms

Stage 1 starts when participants hear the last syllable of the second word of the trigram. Since this syllable is likely to already contain acoustic cues of the

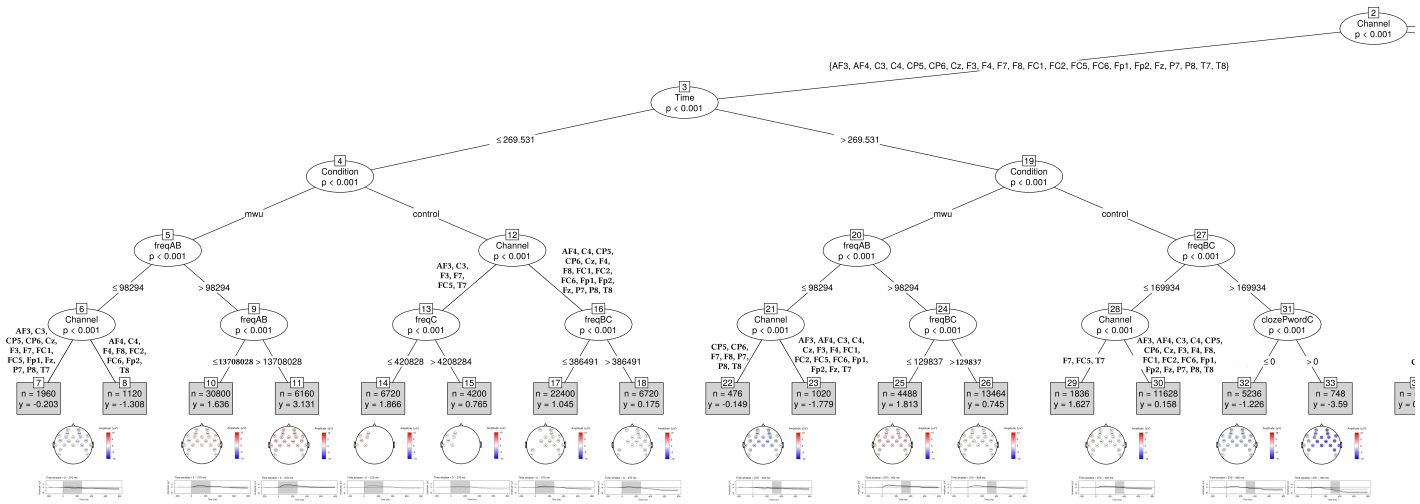

Figure 3.3: Parts of stage 1 and 2 of multi-word processing, visualized here as the first part of a representation of the model-predicted global amplitude values as a function of time, location on the scalp, condition, length of the trigram, unigram, bigram and trigram frequencies, last word cloze probabilities, and word class of the single words included. Global amplitude values are depicted at the terminal nodes, and by means of the colors of the electrodes on the scalps below the terminal nodes. Below the scalps, a time line is provided with the time window concerned highlighted.

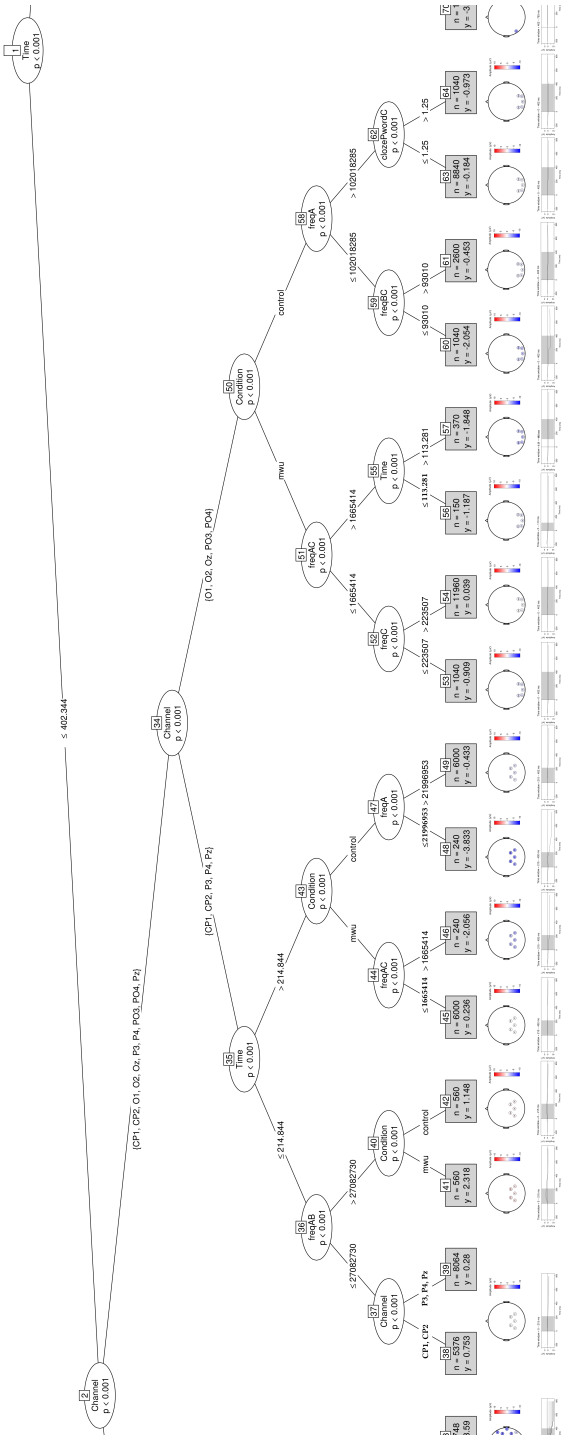

Figure 3.4: Second part of the CForest model, showing what happens at the first and second stages of multi-word unit processing.

following word, we hypothesized that this could be the earliest point in time at which participants notice a difference between multi-word units and matched controls. And indeed, the modeling results already show different neural responses to multi-word units and matched controls between 0 and 240 ms. In short, it appears that participants have already built up expectations on what to expect next, and have different processing strategies for when the bottom-up information matches their expectations (i.e. they encounter the last word of the multi-word unit they were expecting) or when it violates these expectations.

The presence of these different processing strategies suggests that within 270 ms after hearing the first acoustic cues of the last word, participants are sensitive to properties of the full trigram. As can be seen in Figure 3.3, node number three divides a set of frontal, central and temporal electrodes into a time frame before and after 270 ms, and node four subsequently splits this subset by condition, resulting in a subset of multi-word units and a subset of control items. The model has furthermore split the data into different sets of electrodes, which correspond to a fronto-central region, a centro-parietal region, and a parieto-occipital region. In what follows, we will discuss what happens in each of these different regions at stage 1.

Fronto-central processing (nodes 7-18)

At fronto-central regions, processing spoken multi-word units is mostly influenced by the frequency of the first bigram, freqAB, whereas processing of control items is mostly influenced by the frequencies of the last word and the last bigram, freqC and freqBC. We suggest that this reflects different processing strategies when encountering expected or unexpected lexical items, where an expected item prompts the system to further process the first part of the trigram (freqAB), whereas unexpected items prompt the system to shift its attention to these unexpected, new items (freqBC and freqC).

Recall that the ERPs have been time-locked to the onset of the last syllable of the second word, to take into account co-articulatory cues on that syllable. When hearing these cues, participants are able to infer what the last word of the trigram might be. They have already heard and processed most of the first bigram, and have built up expectations as to which word to expect next. Apparently, hearing cues for a word that completes a high-frequency trigram causes participants to continue processing the first part of the trigram. The last part is predictable, and bottom-up processing of the last part is postponed until a later stage. However, upon hearing cues for a word that does not complete a high-frequency trigram, participant's attention is focused towards the last part of the trigram.

When zooming in further into the processing of multi-word units, we see that higher first bigram frequencies are the most important predictor for amplitude values before 270 ms, and that higher AB frequencies correlate with more positive amplitudes. This likely reflects a reduced N1 and an enhanced P2 component. When considering the ERP plots in Figure 3.1, these early com-

ponents are indeed visible, with early onsets and most prominently in bilateral fronto-central regions. In line with our findings, Sereno et al. (1998) found early frequency effects at the N1 and P2 components in a lexicon decision experiment, with high frequency items corresponding to more positive amplitudes. In later studies, Sereno et al. (2003) and Hauk and Pulvermüller (2004) also found frequency effects roughly 150 ms after stimulus onset, with again more frequent items eliciting more positive amplitudes.

For the control items, the model has split the data into two regions: A region in the left hemisphere (nodes 14 and 15) that processes the last word, and a region that is mostly located in the right hemisphere, with electrodes in frontal, central and temporal regions (nodes 17 and 18), that processes the last bigram.

Nodes 14 and 15 represent early processing of the last word. We propose that upon hearing the last word of a control item, participants are prompted to first process this new and unexpected information, before they can integrate it into the previous context. Nodes 14 and 15 show activations in a region that could originate from the left primary auditory cortex (PAC), where auditory processing takes place, or the left inferior frontal gyrus (LIFG), an area that has been connected to linguistic processing in a wide range of studies (see Vigneau et al., 2006, for a meta-analysis of language processing in the left hemisphere). If we follow Friederici's (2012) proposal for a cortical language circuit for auditory processing, then we expect this bottom-up input to pass from the auditory cortex to the anterior superior temporal cortex and then to the prefrontal cortex.

Nodes 17 and 18 of the model represent early processing of the last bigram. The model shows more positive amplitudes for control items with low second bigram (freqBC) frequencies, in mostly right hemisphere regions. The right hemisphere has been implicated in context processing and general attentional and working memory processes (Vigneau et al., 2011) and is claimed to be biased toward bottom-up, more post hoc, interpretive processing (Federmeier, 2007). Considering the relatively large portion of the right frontal hemisphere that is significantly activated in these subsets, it seems plausible that these activations reflect attentional processes and the recruitment of working memory, where the second and third word of the control item are considered at once. Moreover, bilateral peaks in the temporal lobes have been found to be activated during sentence comprehension tasks, and more specifically by tasks where participants had to generate the last word of a sentence (Kircher et al., 2001; Vigneau et al., 2011). As participants were at this point listening to the last part of a trigram, it is likely that they recruit this region associated with sentence completion tasks.

Centro-parietal processing (nodes 38-42)

Before 214 ms, amplitudes at electrodes CP1, CP2, P3, P4 and Pz are more positive for higher first bigram frequencies. When this frequent first bigram is

also part of a multi-word unit, then amplitudes are even more positive. This is similar to the activations seen in the fronto-central regions (see above). Moreover, more positive amplitudes in an early time window for higher frequency items has also been found in previous studies (Hauk and Pulvermüller, 2004; Sereno et al., 1998, 2003).

Parieto-occipital processing (nodes 53-64)

As in fronto-central regions, the most important predictor in this region is condition, which shows that multi-word units are processed differently from controls items in more posterior regions too, and already at an early stage. Note that the terminal nodes of this subgroup are not part of the subgroup of data that has been split into time windows before and after 270 ms — rather, these terminal nodes represent what happens in the centro-parietal regions between 0 and 402 ms after participants heard the first signs of the terminal word of the stimuli. Generally, language processing does not seem to take place in the occipital lobe (Friederici, 2012). However, as EEG is quite imprecise in terms of localization of the neural source, it is possible that the activations seen in the occipital regions originate from more parietal regions.

Multi-word unit processing at parieto-occipital regions is influenced by the frequencies of the last words (freqC) and skipgrams (freqAC), with higher last word frequencies correlating with more positive amplitudes, but with higher skipgram frequencies correlating with more negative amplitudes. More positive amplitudes for higher frequencies also occur in fronto-central and centro-parietal regions. However, it is surprising to see more negative amplitudes elicited by more frequent skipgrams.

The more negative amplitudes for higher skipgram frequencies are unexpected, as the results discussed above all show more positive amplitudes for higher frequencies in multi-word units. Pylkkänen et al. (2004); Tremblay et al. (2016) reported increased activity in their MEG studies around 350 ms (M350) in response to higher lexical and n-gram frequencies. This M350 has been linked to lexical access and indexes inhibitory neural responses. It is possible that high skipgram frequencies cause the listener to consider alternative trigrams, leading to enhanced competition from similar forms, which in turn could lead to increased processing costs as reflected in the more negative amplitudes.

In general, the early posterior activations might reflect the first stage of combinatorial processing and the integration of the last word with the rest of the trigram. In their MEG study investigating language networks involved in n-gram processing, Tremblay et al. (2016) report on a network that is mainly located in posterior regions (their 'Network 3') and whose main function seems to be integrative processing of several sources of information. The network includes areas associated with sentence processing, semantic and discourse coherence processing, and the integration of complex semantic and syntactic information (among others the posterior superior temporal sulcus and the angular gyrus) (Friederici, 2012; Vigneau et al., 2006). As such, it seems probable that

the early activations elicited by last word and skipgram frequencies originate at an integrative network located at posterior areas.

As for control items, we see more negative amplitudes for lower first word (word A) and lower second bigram (bigram BC) frequencies. If the second bigram has a high frequency, however, amplitudes tend to be less negative. This pattern is similar to what we saw in fronto-central regions, where lower frequencies also correlate with more negative amplitudes. Note that we also see an influence of the first word frequency for control items, which we have not seen in more frontal and central regions.

When the first words of a control item has a high frequency, there is also an interaction with the cloze probabilites of the last word of the control item: High cloze probabilities of the last word correlate with more negative amplitudes. This is unexpected, given that this subset of the data is within the time range of the N400 component, and higher cloze probabilities have been found to correlate with a smaller, and thus less negative, N400 component (Kutas and Van Petten, 1994). Moreover, Penolazzi et al. (2007) found more positive-going amplitudes between 280-320 ms at posterior mid-line electrodes for high probability words than for low probability ones.

As might be the case with higher skipgram frequencies correlating with more negative amplitudes, we suspect that the more negative amplitudes for higher cloze probability last words index greater processing costs as a result of inhibitory processes. Once a listener has realized that s/he is not listening to a frequent multi-word unit, s/he is not expecting to hear words with high cloze probabilities, as these are more likely to occur in multi-word units but not in control items[1]. S/he will therefore actively inhibit words with high cloze probabilities. This extra inhibition will make it harder to identify a high cloze probability item.

Discussion Stage 1

In general, in fronto-central regions, low-frequency items elicit more negative amplitudes and high-frequency items elicit more positive amplitudes, with multi-word units eliciting more positive amplitudes overall. The more positive amplitudes of the multi-word units seem to reflect reduced N1 and enhanced P2 components (Sereno et al., 1998, 2003; Hauk and Pulvermüller, 2004). Processing spoken multi-word units is mostly affected by first bigram frequencies, whereas listening to spoken control items is mostly affected by second bigram and last word frequencies.

These differences in processing suggests that listeners engage in predictive processing when they encounter lexical items that could form the beginning of a lexical bundle. If their top-down expectations match the subsequent input, listeners continue processing the first bigram, engaging in lexical selection and

[1]However, this is not necessarily the case. A trigram can have a low phrasal frequency, but a high cloze probability last word. Still, this is less common and therefore less expected.

perhaps also lexical integration of that first bigram. However, if their expectations do not match the bottom-up input, listeners focus their attention towards the unexpected input and start processing the last word and the last bigram, engaging in the first stage of spoken word recognition, lexical access.

Posterior regions are also involved at this stage, and even seem to be engaged in later processes of spoken word recognition, i.e. combinatorial and integrative processing. These regions are involved in the processing of skipgrams and the last words of multi-word units, and the first words of control items. Moreover, the cloze probability of the last word of control items also plays a role in these regions. Although the activations could have originated from the primary auditory cortex and therefore only reflect the first stage of auditory processes, it seems likewise plausible that the activations also reflect the involvement of the posterior superior temporal sulcus and the angular gyrus, which perform combinatorial and integrative processing (Friederici, 2012; Vigneau et al., 2006).

It is not only because of the possible neural source that we suspect that the activations seen in posterior regions index later stages in spoken word recognition; it is also because of patterns of activations we observe. Higher-frequency items that correlate with more negative amplitudes likely reflect higher processing costs. Different processes seem to underlie these enhanced processing costs: For higher frequency skipgrams, the larger costs are likely a result of lexical competition effects, and reflect a form of neighborhood density effects (Luce and Pisoni, 1998). For higher cloze probability words (words C of control items), the larger costs are likely a result of inhibitory effects. Most control items are likely to end in a low cloze probability word, and as soon as a listener is aware that s/he is listening to a control item, s/he seems to actively inhibit items that s/he is not expecting to hear, i.e. words with a high cloze probability.

3.4.3 Stage 2: Inhibition and competition, 270-402 ms

The most important data split in stage 2 is type of n-gram: Multi-word unit processing is influenced mostly by the first bigram frequencies (freqAB), whereas control item processing is influenced mostly by the second bigram frequencies (freqBC). In stage 1 there are already some forms of inhibitory and competitory processing in posterior regions. These processes continue and become more prominent and widespread in stage 2. Higher frequency second bigrams in both controls items and multi-word units elicit more negative amplitudes, reflecting larger processing costs: In multi-word units high frequency second bigrams seem to prompt processes of competition between similar forms, whereas high frequency second bigrams in control items are unexpected and prompt processes of inhibition.

Fronto-temporal processing (nodes 21-33)

For multi-word units, lower AB frequencies correlate with more negative amplitudes at fronto-temporal locations between 270 and 402 ms, similar to the

direction of the effect before 270 ms. Hagoort and Brown (2000) found a large negative shift around 250 ms for semantically anomalous words, with a mostly central distribution. The authors hypothesize that the N250 might reflect the lexical selection process that takes place at the interface of lexical form and context integration. Although having lower frequencies is not the same as having a semantically anomalous form, the phenomena are similar in that they constitute less expected events.

For multi-word units with high AB frequencies, the frequencies of the second bigram also start to play a role. The higher the first bigram frequency is, the faster participants will be in processing that trigram, which results in earlier onsets of the processing of its second bigram. High BC frequencies, however, correlate with less positive amplitudes than low second bigram frequencies (nodes 25 and 26). Higher first bigram frequencies elicit more positive amplitudes, whereas higher second bigram frequencies elicit less positive amplitudes. This is similar to the more negative-going amplitudes for high frequency skipgrams, as we saw in the subsection on parieto-occipital processing in Section 3.4.2. Therefore, we suspect that the less positive amplitudes elicited by higher second bigram frequencies in this time window index competitory processes (Pylkkänen et al., 2004; Tremblay et al., 2016).

When considering the control items, there is also a negative correlation between frequencies and amplitudes: Both higher BC frequencies and higher cloze probabilities of the last word correlate with more negative amplitudes (node 33). Once a listener has arrived at stage 2 of processing, s/he has already realized s/he is not listening to an expected multi-word unit, and therefore expects a low-frequency second bigram and a low cloze probability item as the third word. S/he might be actively inhibiting high-frequency bigrams and high cloze probability third words, which could lead to enhanced processing costs for these parts of control items, reflected as more negative amplitudes.

Centro-parietal processing (nodes 45-49)

Between 214 and 402 ms at centro-parietal regions, amplitudes elicited by multi-word units are mostly influenced by skipgram frequencies (freqAC; nodes 45 and 46). Higher skipgram frequencies correlate with more negative amplitudes for multi-word units, possibly reflecting larger processing costs due to enhanced competition from similar forms (Pylkkänen et al., 2004; Tremblay et al., 2016). This is similar to what happens between 0 and 402 ms in parieto-occipital regions (see Section 3.4.2). It seems then, that these competitory processes originate in posterior regions and move forward to (or happen concurrently in) more centro-parietal regions. Control items, on the other hand, are mostly influenced by the frequencies of the first word (nodes 48 and 49), with higher first word frequencies correlating with more positive amplitudes.

Parieto-occipital processing (nodes 53-64)

See 'Occipital-parietal processing' in Section 3.4.2.

Discussion Stage 2

In stage 2, bigrams and words of multi-word units are further processed and integrated. Multi-word unit processing is influenced by the first bigram frequencies and, to a lesser extent, by the second bigram frequencies in frontal and central regions, and by skipgram frequencies (freqAC) in more centro-parietal regions. Processing of control items is influenced by the second bigram frequencies and the cloze probability of the last word in fronto-central regions, and by the frequencies of the first word in centro-parietal regions.

Like in stage 1, multi-word units elicit more positive amplitudes overall. These positive amplitudes seem to reflect a reduced N250 (Hagoort and Brown, 2000). The occurrence of an N250 suggests that at this point in time, lexical selection of the target trigram takes place if the trigram is a frequent multi-word unit. However, if the trigram is a low-frequency control item, then the language system spends more resources on processing the last part of the trigram.

By now, the first signs of the influence of the last words of high-frequency multi-word units start to appear. In fronto-central regions, higher second bigram frequencies elicit more negative amplitudes. These more negative amplitudes seem to reflect enhanced processing costs, which are likely due to competition effects between similar high-frequency bigrams that could complete the first frequent bigram. Similarly, in centro-posterior and in occipital regions, higher skipgram frequencies of multi-word units elicit more negative amplitudes, indexing competition effects of similar skipgrams.

These reflections of competition effects for both bigrams and skipgrams might originate from the angular gyrus and the posterior superior temporal sulcus, locations which have been connected to syntactic and semantic integration and sentence processing tasks (Tremblay et al., 2016; Vigneau et al., 2006). They have moreover been linked to semantic processes at the sentential level (Lau et al., 2008). Therefore, we propose that the emerging activations of the last bigram and skipgram frequencies in these regions reflect integrative processes that link the beginning of the multi-word unit to its ending.

As for control items, we see a continued influence of the last word's cloze probability. Higher cloze probability items are more unexpected in low-frequency trigrams, and as such, elicit more negative amplitudes. Moreover, higher last bigram frequencies also elicit more negative amplitudes, again because the presence of a high-frequency last bigram is unexpected in a low-frequency trigram, which increases processing costs.

3.4.4 Stage 3: Lexical integration, 402-800 ms

At the third stage, the most important split in the data is again made by condition: 402 ms after hearing the first signs of the last word of a trigram,

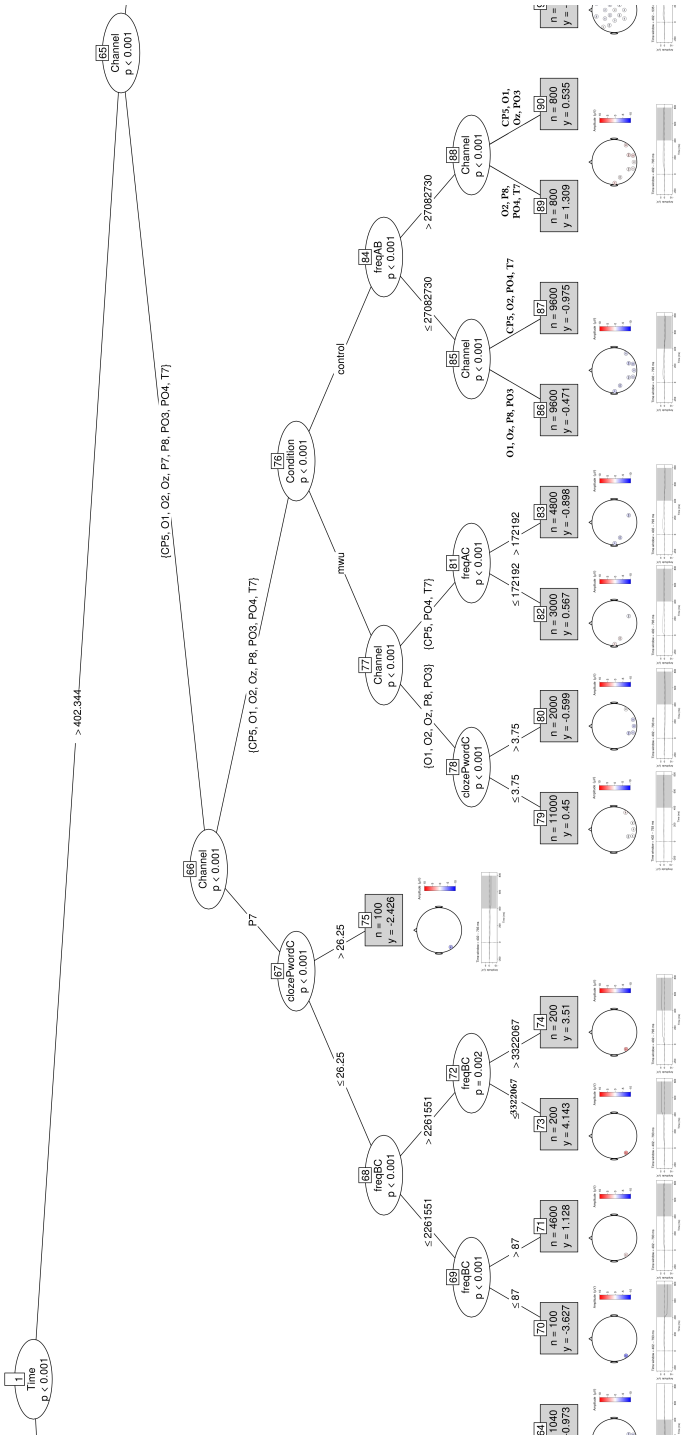

Figure 3.5: Third part of the CForest model, showing what happens at the third stage of multi-word unit processing.

Figure 3.6: Fourth part of the CForest model, showing what happens at the third stage of multi-word unit processing.

frequent multi-word units are still processed differently than matched control items. Specifically, there is a sustained negativity that is less negative for multi-word units and most pronounced in central and right-hemispheric electrodes in fronto-central regions (see also Figure 3.1).

Fronto-central processing (nodes 95-121)

At fronto-central regions, multi-word units are first split by time, into a time window from 402 - 535 ms, and a time window after 535 ms. Both time windows are mostly influenced by the frequencies of the last word, with higher last word frequencies correlating with more negative amplitudes. When the last word frequencies are low, but the last bigram frequencies (freqBC) are high, then amplitudes are even more negative (node 96).

Given that previous research has shown that, for lexical access of single words, lexical integration is taking place after 400 ms (and possibly sooner; Steinhauer et al., 2008), and since we see more negative amplitudes after 400 ms for higher frequency items, we propose that negative amplitudes are indications of easier lexical and contextual integration in stage 3 (Friederici, 2012). This is in contrast to what happens at stage 1 and 2, where more negative amplitudes seem to reflect processes of competition or inhibition.

More negative amplitudes that index ease of processing at this stage 3 are likely to be reduced P600 components. The P600 is an ERP component which is typically elicited by grammatically erroneous sentences, with the incorrect sentence eliciting a greater positivity compared to the correct one. It has been reported to surface as early as 400 ms after stimulus onset (Kaan and Swaab, 2003). Besides grammaticality, the P600 has been reported to vary according to the effort needed to build a coherent syntactic structure (Hagoort, 2003), to reflect continued combinatorial analyses efforts of the brain (Kuperberg, 2007), and to vary according to the degree of probability and salience of a sentence, with more probable sentences eliciting a reduced P600 (Coulson et al., 1998). Frequent BC bigrams and last words of multi-word units are expected, probable, and should therefore take up less processing effort, which would then translate in a less positive, i.e. a more negative amplitude.

In contrast, control items are less likely to be processed as chunks, which means that more combinatorial processes must be at work for control items (Kuperberg, 2007), increasing the P600, leading to more positive amplitudes for control items overall. The most important predictor of amplitude values in these control items is the frequency of the second bigram. Like for the multi-word units, higher frequencies and higher probabilities seem to reduce a P600(-like) component (Coulson et al., 1998). Generally, the higher the second bigram frequency, the more negative the amplitude. If moreover the cloze probability of the last word is also high, then amplitudes are even more negative (node 120 and 121). This reduction in a positive component can also be seen for control items with a low cloze probability last word, as long as their first word has a high frequency; when the first word has a low frequency, amplitudes are more

positive (nodes 117 and 118).

When the last bigram of control items is not frequent, amplitudes vary mostly in terms of region on the scalp (node 108). In the left hemisphere, the word class of the second word matters most, with open class words eliciting a more negative amplitude. This more negative amplitude probably reflects ease of processing, as discussed above. As over 90% of the first words of the trigrams start with a closed-class word, most participants will have expected the second word to be an open-class word[2]. When this expectation is violated, a more positive amplitude is elicited (node 110), resembling a P600 effect elicited by unexpected events as discussed by Coulson et al. (1998). In the right hemisphere and central locations, amplitudes vary over time, where amplitudes after 543 ms are more negative. It is likely that these amplitudes reflect context processing, which is known to happen in the right hemisphere (Vigneau et al., 2011).

Parieto-occipital processing (nodes 70-90)

Node 66 separates electrode P7 from the other parieto-occipital electrodes. A low cloze probability of the last word leads to more positive amplitudes at this electrode than high cloze probabilities of the last word (node 75), again showing that, at this third stage, unexpected or improbable events elicit more positive amplitudes. Interestingly, when the cloze probability of the last word is low, and when the frequency of the last bigram is also low, amplitudes are much more negative (node 70). It is not clear why only electrode P7 is split from the other subset of electrodes, and it might be the case that the model is overfitting the data. Note, moreover, that this bin only contains 100 data points, making it unlikely that this effect is robust and generalizable. Future studies could ascertain whether or not this effect is robust.

For the other parieto-occipital electrodes, amplitudes vary by condition. The cloze probability of the last word plays a role in multi-word unit processing in a region of electrodes O1, Oz, O2, P8 and PO3, whereas the skipgram frequencies play a role in multi-word unit processing in a region of electrodes CP5, PO4, and T7. Especially this last region is surprising, as it constitutes a non-continuous region in both the left and right hemisphere. As participants had to judge, at random intervals, whether or not a visually presented fragment could be a correct continuation of the stimuli presented to them, it is possible that this is a prediction network where the skipgram is aiding the lexico-semantic system (CP5, T7) in suggesting possible continuations, which in turn feeds into the visual cortex (PO4) to prepare for a possible visual stimulus. Increased activations in the visual cortex indexing the pre-activation of predicted visual features were also found by Dikker and Pylkkänen (2013).

For control items (nodes 86-90), it is mostly the frequency of the first bigram that plays a role, with higher first bigram frequencies correlating with more

[2]In the top 1,000 trigrams from the TenTenCorpus (Jakubíček et al., 2013), out of which our stimuli have been sampled, 81.6% of these frequent trigrams start with a function word. It seems then, that frequent trigrams tend to start with function words in English.

positive amplitudes. High frequency first bigrams are unexpected in control items, which might explain why more unexpected items elicit more positive amplitudes here, indexing a larger P600 response (Coulson et al., 1998).

Discussion Stage 3

Given the regions involved and the presence of a P600 component, it is probable that during stage 3 lexical integration of all elements of both multi-word units and control items takes place. Moreover, the frequencies of the BC bigrams are playing a clear role in the processing of multi-word units, showing that at this point the last part of the trigrams is also processed and integrated. Whereas higher frequencies correspond to more positive amplitudes for multi-word units in the first two stages, higher frequencies correspond to more negative amplitudes in the third stage. This, we proposed, is likely to be a reflection of a reduced P600 component indexing ease of lexical integration.

The previous two stages involved more positive amplitudes for items that are easier to process. However, at this stage, more negative amplitudes are indicators of ease of processing. A likely ERP component for this stage that reflects ease of processing is a reduced P600. As multi-word units have a higher phrasal frequency, are more expected, and do not necessarily need combinatorial processes, a reduced P600 response is not unexpected. Moreover, higher frequency single words, bigrams, or higher cloze probabilities of items in both multi-word units and control stimuli are also more probable, easier to process, and therefore more likely to elicit reduced P600s — which manifests itself in more negative amplitudes.

3.5 General discussion

We have collected ERP data of participants listening to both multi-word units and their matched controls. We selected a group of high-frequency trigrams and created a set of matched controls by changing the last word for another word that was just as frequent as the original word, but that would not form a frequent combination with the first two words. This way, we could compare processing of high-frequency trigrams and low-frequency similar trigrams that only differed in their last parts.

When a listener encounters a stream of words, at first, s/he cannot know if s/he is listening to a multi-word unit, a low-frequency combination of words, or even a meaningless combination of random words. Because listeners constantly update their expectations based on what they encounter in this stream of words, we expect them to also form expectations on whether they are listening to the first part of a multi-word unit or a low-frequency combination of words [3]. If listeners have different expectations on what to hear next, we expect them to

[3]They will likely not expect a combination of random words, as verbal communication typically carries a meaning and a message.

also employ different processing strategies after hearing at least the first word of a trigram, which should also manifest as different ERP patterns. In other words, we expect to see differences in the ERP data at the moment listeners have already heard and partly processed the first part of a multi-word unit. So to understand if and how spoken multi-word units and matched controls are processed differently, we focussed on the processes taken place after a listener has already listened to the first word and (a part of) the second word of a trigram.

First of all, the ERPs show a clear difference with an early onset between the two conditions. This provides clear evidence for the expectations formulated above, i.e. that spoken multi-word units and matched controls are processed differently. This must be due to the frequency of the combination, as the individual words were matched for their individual frequencies.

Secondly, the different ERPs and their different manifestations provide indications as to what the nature of these differences is: Ease of processing. Overall, we found a sustained negativity that is more positive for multi-word units. Multi-word units show reduced N1 and P2 components, a reduced N250, and a reduced P600 as compared to control items. All these features suggest that multi-word units are easier to process than non-frequent combinations of words (Coulson et al., 1998; Sereno et al., 1998; Hagoort and Brown, 2000; Hagoort, 2003; Sereno et al., 2003; Hauk and Pulvermüller, 2004; Kuperberg, 2007).

Previous experimental work has already shown that ease of processing is manifested as an increase in speed in naming (see e.g. Arnon and Snider, 2010), and greater accuracy in recall (Bannard and Matthews, 2008). In this study, we did not find faster listening per se, but different processing strategies in how multi-word units are processed in comparison to control items. For future studies it will be interesting to explore the possibility that easy of processing in the case of listening to a multi-word unit is also manifested as greater accuracy in processing the auditory signal.

Thirdly and finally, by studying parts of a CForest model, we were able to come up with a detailed proposal on how auditory processing of multi-word units and their matched controls might proceed, and which factors contribute most. For this study, we only focused on the time window where the auditory signal of multi-word units and their matched controls starts to diverge, i.e. from the last syllable of the second word onwards. In view of the early onset of the differences between the conditions in this study, it would be informative to also consider the processing of the first part of spoken trigrams, thereby studying the full course of processing of whole multi-word units.

Our analysis suggests that there are three stages in time during which the last part of either a frequent multi-word unit or a matched control item is processed. The first stage consists mainly of predictive and bottom-up processes, where more positive amplitudes indicate ease of processing. The second stage revolves around combinatorial processes that are influenced by competitive and inhibitory processes, and where again more positive amplitudes indicate ease

of processing. The third and final stage consists of integrative processes, where more *negative* amplitudes indicate ease of processing. Units from different levels of complexity play a role in processing, with trigram, bigram, unigram frequencies, and word types of single words playing a role concurrently or in close approximation in time or location. Similar results were found by Tremblay and Baayen (2010), who also found that quadgram probabilities as well as sequence-internal word and trigram frequencies affected event-related potentials.

One of our key proposals is that listeners adapt their processing strategy on the basis of what they expect to hear and what they actually hear — at first focusing more on the first part when hearing a multi-word unit, but more on the last part when listening to a control item — which shows an influence of top-down processes on further processing. These different processing strategies offer evidence in favor of interactive models of auditory processing, where multiple sources of information are employed in parallel (Brink and Hagoort, 2004; Hagoort, 2003; Tremblay et al., 2016).

Acknowledgements We would like to thank Daan van de Velde for his help in creating the stimuli.

CHAPTER 4

Reading and speaking

Keeping it apart: on using a discriminative approach to study the nature and processing of multi-word units

Saskia E. Lensink, Arie Verhagen, Niels O. Schiller, R. Harald Baayen

abstract
A growing number of studies finds frequency effects for common combinations of words, leading many to assume that these multi-word units have some kind of cognitive reality. However, it is not clear how lexical access to these multi-word units takes place. We conducted two experiments, where the tracked the eye movements and recorded the voices of participants reading silently and out loud through a list of frequent multi-word units, and modeled the data using both traditional measures of lexical access and measures taken from a computational model of lexical access that incorporates multi-word units, the Naive Discriminative Learner (NDL). Results show that the NDL measures provide additional insights, showing that lexical access to multi-word units proceeds from top-down to bottom-up processes, with larger co-activations of similar items speeding up production. Moreover, the eye-tracking data shows that readers are faster in reading multi-word units when they spend more time at initial stage of reading, i.e. the first pass.

Keywords: word naming, eye-tracking, multi-word units, phrasal frequency effects, naive discriminative learning, Rescorla-Wagner equations

4.1 Introduction

A large part of language is formulaic in nature. Common combinations of words are claimed to make up at least twenty percent of total usage in spoken and written language (Erman and Warren, 2000). A growing number of experimental studies has reported frequency effects for combinations of two or more words (Arnon and Snider, 2010; Shaoul and Westbury, 2011, and references therein). Several studies have looked at frequent multi-word units in both production and comprehension, using experimental paradigms such as self-paced reading, phrasal decision tasks, and word reading tasks. Moreover, different techniques have been used, including EEG and eye-tracking (Siyanova-Chanturia, 2013). Most work has focused on multi-word unit processing in adult native speakers, but several studies also consider processing in children (Bannard and Matthews, 2008) and L2 speakers (Conklin and Schmitt, 2012; Han, 2015; Jiang and Nekrasova, 2007; Siyanova-Chanturia et al., 2011b).

Although there are some differences between the findings of these studies, an overall finding that emerges consistently is an effect of the frequencies of multi-word units, even when the frequencies of the individual words have been

controlled for. The phrasal frequency effect has been interpreted as evidence for "holistic" multi-word units in the mental lexicon, or as evidence for experience in using the rules of grammar supporting these multi-word units (Arnon and Priva, 2014; Siyanova-Chanturia, 2015; Tremblay et al., 2011).

Considering this previous research, there is abundant evidence that multi-word units play a role in processing. The question of how, given some input, a lexical unit is accessed is central to all models addressing language comprehension and production. However, we know very little nor do we understand how lexical access to multi-word units proceeds. This study aims to fill this gap by investigating the lexical access of multi-word units by means of combining a computational modeling study with newly collected experimental data. The computational model of choice is a Naive Discriminative Learning network (NDL; Baayen et al., 2011); the data are collected in an eye-tracking study and a reading aloud study.

4.1.1 Including multi-word units in models of lexical access

Previous research has shown that frequency effects for multi-word units could be predicted by a model that did not have any representations for multi-word units itself (Baayen et al., 2013). The phrasal frequency effect was merely an emergent property of a network that implemented error-driven learning, crucially without specifying any phrasal units.

The reason for not implementing these units was that there are several drawbacks to the idea of storing multi-word units in the mental lexicon. One such drawback is that there are hundreds of millions of word n-grams that would need to be stored (Baayen et al., 2013), even under the assumption that n is unlikely to be much larger than five or six (Shaoul et al., 2013, 2014a). Populating the mental lexicon with such vast numbers of representations raises issues not only of storage, but also of increased retrieval costs.

So why still consider including full multi-word units in models of lexical access given these drawbacks? We may be underestimating the memory capacity of our brain. We have a vast inventory of detailed experiences of the world stored in our memory (see e.g. Brady et al., 2008). Storage of our experience with language is likewise huge. Not only do we store information about the meanings of words, but also about the different phrasal contexts in which these words can be used and the different meanings connected to these contexts, pragmatics, as well as different syntactic constructions and their meanings, to name just a few. Baayen et al. (2011) and Milin et al. (2009) have shown that inflectional, derivational and even prepositional paradigms play a role in language processing, suggesting we store all this information. Furthermore, recent research on Estonian, a Finno-Ugric language related to Finnish, documents form frequency effects for case-inflected nouns (Lõo et al., 2017, 2018), in this language the functional equivalent of prepositional phrases in English.

Given the vast knowledge we have of the world, and of language, the reflection of this knowledge in language processing — in the form of a phrasal frequency effect — should perhaps not be surprising. Moreover, when we consider the stimuli chosen in many of the experiments studying phrasal frequency effects, it transpires that many of the multi-word units used encode relevant and meaningful experiences. These units concern very specific time markers, such as 'on the day', discourse markers such as 'I think that', and affordance relations, such as 'on the table'. These experiences are easily conceptualized as being united and therefore as single units of experience.

Although conceptually and referentially transparent (unlike idioms), these multi-word units have properties that are distinct from the sum of their parts, which must be represented somewhere and are expected to play a role in processing. It seems likely that single words, idioms, and certain multi-word units are essentially the same type of entity psychologically. This is reminiscent of one of the central tenets of constructionist approaches, where there is no principled difference between morphemes, words, and constructions (Bybee, 2010; Croft, 2001; Goldberg, 2003). Therefore, there are good reasons to treat at least some multi-word units in the same way as single words (Baayen et al., 2011) or idioms (Geeraert et al., 2017).

To summarize, there are both empirical and theoretical reasons to take multi-word units into account in our models of lexical access. Experimental evidence has shown that they influence processing, and that it is plausible that we store a lot of forms, given our huge storage capacity. Furthermore, several frequent combinations of words encode experiences separate from the sum of their parts, which could results in the creation of unitary multi-word time markers, discourse markers, and affordance relations.

4.1.2 Computational modeling of multi-word units

To explain previous findings of phrasal frequency effects, it is not enough to only consider the frequency with which language users are exposed to multi-unit words (Baayen, 2010). We also need to know to what extent the smaller parts of a multi-word unit form informative cues to access the full multi-word unit and how language users are able to keep different multi-word units apart. We take a discriminative learning approach, using a computational model that incorporates principles of learning theory (Baayen and Ramscar, 2015; Baayen et al., 2011; Ramscar and Yarlett, 2007; Ramscar et al., 2010) using the Rescorla-Wagner equations (Rescorla et al., 1972).

The model of choice, Naive Discriminative Learning (NDL; Baayen et al., 2011), has several advantages: first, we understand the inner workings of the model quite well as it consists of only two layers; second, NDL models provide us with measures that show how lexical access could proceed (**?**); third, it is a cognitively plausible model as it incorporates principles of learning theory, which we believe are essential in understanding how language works (see e.g. Baayen and Ramscar, 2015; Arnon and Ramscar, 2012); fourth, the model

scales up to large lexicons (Arnold et al., 2017); fifth, software to implement this model in R or Python is freely available (R: `ndl2`; Shaoul et al., 2014b), python: available at `github.com/quantling/pyndl`); sixth, and relevant for this study, NDL allows for a straightforward implementation of multi-word units.

For this study, we did not make use of other models of lexical access, as there are no viable alternatives that allow us to understand lexical access to multi-word units. TRACE (McClelland and Elman, 1986) does not scale up to large lexicons, and it is not clear how to implement multi-word units in the model. The same limitations apply to the Shortlist-B model (Norris and McQueen, 2008).

In what follows, we will discuss how NDL models function in general, and how we have implemented multi-word units in an NDL model. We will then present new experimental data on reading and producing common Dutch multi-word units, and will test to what extent the NDL measures add anything over and above the more traditional frequency measures in modeling this data. We conclude with a discussion of what our findings tell us about how lexical access to multi-word units proceeds.

4.2 NDL model

Learning is not just the result of keeping track of how often a certain cue predicts an outcome. It is also dependent on how informative a cue is in light of other cues that predict the same outcome, and in light of other outcomes that are predicted by that cue. These aspects of learning can be captured by the learning equations developed by Rescorla et al. (1972), which are closely related to the learning rule of Widrow and Hoff (1960) and the perceptron (Rosenblatt, 1958). These equations do not only predict animal behavior, but are also able to predict aspects of implicit learning (Ramscar et al., 2010, 2013; Ramscar and Yarlett, 2007).

Recently, Baayen et al. (2011) implemented the Rescorla-Wagner equations in a computational model for language learning: naive discriminative learning (NDL). NDL networks have been shown to predict a wide range of linguistic phenomena such as lexical decision latencies, word frequency effects, phrasal frequency effects, and ERP amplitudes. Its predictions are moreover consistent with the performance of young infants in an auditory comprehension task (Baayen et al., 2011; Baayen and Ramscar, 2015). For technical details we refer the reader to Baayen et al. (2011).

4.2.1 How the model works

We will briefly describe how the NDL network works conceptually. An NDL network consists of only two layers: a layer of input units (henceforth cues) and a layer of output units (henceforth outcomes). By implementing this network

we obtain a mathematical characterization of how well outcomes can be discriminated given some set of input cues. Since the weights to outcome i are estimated independently from the weights to outcome j, the model is "naive" in the sense that it does not exploit information about how outcomes co-occur.

Cues can be formed by low-level features, often letter bigraphs or trigraphs, or single words, like we did in this study. Outcomes are formed by pointers to a location in a high-dimensional semantic vector space (see Landauer and Dumais, 1997; Lund and Burgess, 1996; Shaoul and Westbury, 2010; Mikolov et al., 2013, for detailed discussion of such models). This location can reflect a single word, a grammatical feature, an idiom (Geeraert et al., 2017), or a multi-word unit. To clarify that the outcomes in an NDL model are not units of form, nor monadic "meanings", but pointers to semantic vectors, these pointers are called *lexomes* (Milin et al., 2017; Baayen et al., 2017b). They are best understood as stable mediators between variable linguistic forms - the cues - and variable experiences of the world.

In an NDL model, every cue is connected to all outcomes and every outcome is connected to all cues. The weights of these connections are estimated from a corpus. As a first step, learning events have to be derived from the corpus. A learning event is defined by a set of cues and one or more outcomes that are jointly evaluated by the Rescorla-Wagner learning rule. Learning events can comprise single words (see, e.g., Arnold et al., 2017; Linke et al., 2017), or multiple words (cf. Baayen et al., 2011, 2017b; Geeraert et al., 2017).

The model learns by going over sentences of a corpus one by one, updating the weights from cues to outcomes, based on the information present in that specific learning event. At each step, the predictions of the network given the cues in the learning event are compared with the outcomes in the learning event. When a cue and an outcome are both present, their association weight is strengthened. Conversely, when a cue occurs without an outcome, their association weight is weakened.

A cue is informative and thus discriminative if strong connection weights lead to only a small number of outcomes. However, if a cue is more or less evenly connected to a lot of different outcomes, then this specific cue cannot be a strong predictor of any of the outcomes. Articles are bad predictors of the identity of any multi-word unit, for example, whereas the word *happily* is a strong discriminative cue for the outcome *happily ever after*.

For the modeling of lexical access to multi-word units, we specified learning events and the cues and outcomes therein. As learning events, we used the 19,091,130 utterances in a Dutch subtitle corpus, which comprises 109,807,716 word tokens. Since our working hypothesis is that multi-word units are cognitive units, the outcomes of the network will represent such units. We selected a set of 296 trigrams - combinations of three words - that frequently occur in the Dutch language and that have a transparent meaning. A transparent trigram does not have a figurative or idiomatic meaning; the meaning of the whole trigram can be deduced from the sum of the meaning of its parts. *On the table*

is an example of such a transparent trigram.[1]

The question now arises what outcomes for multi-word units might represent. Given that in naive discriminative learning the outcomes represent pointers to semantic vectors, we propose to interpret the outcomes for multi-word units in the same way. Interestingly, in semantic vector spaces, operations can be defined such that the semantic vector for one word, e.g., *sister*, is a mathematical function of the semantic vectors of related words, e.g., *female* plus *sibling* (see, e.g., Mitchell and Lapata, 2008; Mikolov et al., 2013; Lazaridou et al., 2013). Therefore, the lexome for a word trigram such as *the president of* could likewise be a pointer to a location in the semantic space that is some compositional function of its constituents. Unlike the case of *female sibling*, where a separate word co-exists (*sister*), Dutch and English multi-word units have no such single-word counterpart. However, note that there are language where such meanings as *the president of* are encoded as a single word.

Furthermore, multi-word units could highlight different perspectives on or affordances of objects or actions. For instance, the trigram *the president of* may highlight that presidents are officers having responsibilities for and power over countries or organizations, whereas *president* in an utterance such as *Mr. President* functions as a title and formal mode of address.

As input units, we defined the cues as all the unique individual words in the utterance. This model set-up stays close to approaches in which higher-level units are predicted primarily from the units one level lower in a hierarchy of units for ever smaller features.

So our NDL model used in this study takes single words as its input cues, and multi-word unit lexomes as its outcomes. We made use of the `ndl2` package for R (Shaoul et al., 2014b), which runs on linux only. A platform-independent implementation in python is available at `github.com/quantling/pyndl`. The learning rate (the product of the α and β parameters in the Rescora-Wagner equations) was set at 0.001, and the λ parameter (representing the maximum evidence) was set at 1.0. See Figure 4.1 for a graphical representation.

4.2.2 NDL measures of lexical access

From the model we can calculate several different measures that reflect the availability of trigrams, bottom-up activation of trigrams, and the uncertainty about the identity of trigrams. These measures have been found to be strong predictors of lexical processing.

The first measure is the L1-norm of an outcome, henceforth the outcome's *prior*. It is calculated by summing over all the absolute values of the afferent weights that lead to a specific trigram.The L1-norm is a distance measure. It can be understood as the distance covered when a point can be reached only by traveling along one of the axes at a time. Thus, in the two-dimensional

[1]Still, despite their transparent meaning, we do suspect that frequent multi-word units do encode additional meanings in that they often function as time or discourse markers, and affordance relations.

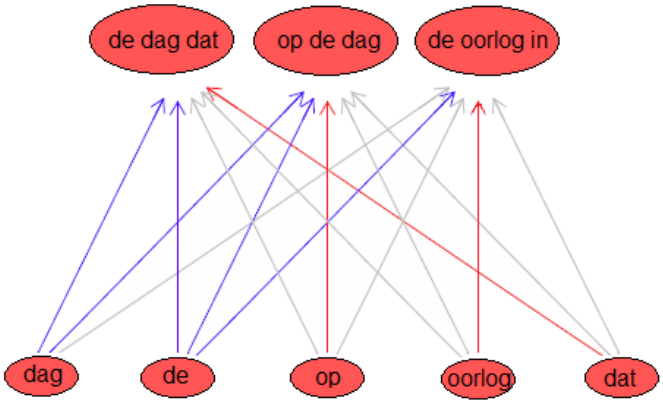

Figure 4.1: Part of the Rescorla-Wagner network used in this study, where the cues are formed by single words and the outcomes by the word trigrams used in the two experimental studies. Each cue is connected to all outcomes, and vice versa all outcomes are connected to each and every single cue. Red lines indicate strong support for a certain multi-word unit, blue lines a weaker support and the grey lines very weak support.The Dutch *de dag dat*, *op de dag* and *de oorlog in* mean "the day that", "on the day" and "into the war", respectively.

plane, the distance traveled to reach the point (3, -4) is 3 units along the horizontal axis plus 4 units along the vertical axis, a total of 7 units. (The L2-norm of a vector is the more familiar Euclidian distance, the distance covered when traveling straight from the origin to the point (3, -4), thus, the Euclidian distance is 5.) Assuming that the groups of neurons underlying cues have a background firing rate - seen in several kinds of neurons - the prior reflects how active an outcome is when there is no visual input. In other words, this L1-norm provides a measure of network entrenchment that is independent from the input and functions as a proxy for resting-state activity. For detailed discussion of this measure, as well as empirical evidence for its predictivity for lexical processing, see Milin et al. (2017).

The prior is strongly correlated with frequency of occurrence in the corpus on which the network is trained. Indeed, the correlation between NDL priors and trigram frequencies for our data is as high as 0.96. We systematically explored which of the two measures performed the best, and kept only the predictor that explains most of the variance in the data. In some of our models, the frequency predictors performed slightly better, in other models the NDL priors. We expect that higher frequencies and priors will lead to shorter fixation durations or a lower number of fixations in our eye-tracking data, and shorter production durations in our production data.

The second measure taken from the NDL networks, the *activation* of an outcome unit, is the sum of the weights on the connections from the cues that are present in the input to that outcome. This measure gauges the bottom-up support for an outcome. Activations are predictive of a wide range of linguistic phenomena, such as lexical decision latencies, word frequency effects, phrasal frequency effects, and ERP amplitudes (Baayen et al., 2011; Baayen and Ramscar, 2015; Hendrix et al., 2017; Baayen et al., 2016a). Higher activations indicate easier processing. Therefore, we expect that in our data higher activations will correlate with either shorter fixation durations or a lower number of fixations, and shorter production durations.

The third measure, the *activation diversity*, gives an indication of the uncertainty regarding the identity of a trigram. It assesses the extent to which activation is dispersed over many different outcomes with the L1-norm of the efferent weights of the cues in the input to all outcomes. The larger the activation diversity is, the larger the number of other outcomes that are also supported by the cues in the input. One can think of this measure as quantifying the extent to which the cues perturb the distribution of the outcomes' priors. In an ideal situation, the cues in the input would support only the targeted outcome, leaving all other outcomes completely unaffected. In such a case, the perturbation of the priors of the outcomes would be minimal. However, in reality, learning is seldom this crisp and clear-cut, and the states of outcomes other than the targeted ones are almost always affected as well, sometimes substantially. The more the distribution is perturbed, the greater the uncertainty about which outcome is the targeted outcome. Conceptually, the activation diversity resembles measures of neighborhood density.

Slower latencies are expected when the diversity values are high, as higher values indicate that many other irrelevant outcomes are also highly activated. Indeed, Milin et al. (2017) found slower response latencies for increasing values of activation diversity in lexical decision experiments. Likewise, Arnold et al. (2017) found that higher activation diversity values correlated with longer latencies in auditory lexical decision. We expect that in our eye-tracking data, high diversity values will correlate with longer fixation durations or more fixations, and in our production data, that high diversity values will correlate with longer production durations.

One technical note is in order with respect to how we estimated activation diversities. Because NDL implements *naive* discrimination learning, it is not necessary (even if it were possible) to include huge numbers of word trigrams in the simulation. Because the weights on the connections from the cues (25,163 letter triplets) to a given outcome are estimated independently for each outcome, it suffices to include in the simulation only the 296 word trigrams used in the experiments below. For each learning event in which none of the 296 word trigrams were present, a dummy trigram was included as outcome. This ensures that weights on connections from cues to the target trigrams are properly decreased across all learning events. Activation diversity for a given set of input cues is calculated over the vector of activations over all trigrams, including the dummy, that these cues give rise to.

4.3 Generalized additive mixed models

How exactly the three NDL measures work together to predict an experimental response variable is not specified by NDL theory. In general, higher activations and priors should reflect reduced processing costs, whereas a higher activation diversity should predict increased processing costs. But whether they interact, and if so, how, is not straightforwardly predictable. As a consequence, models using discrimination measures are intrinsically exploratory in nature, and we will depend on generalized additive mixed-effects modeling to screen the data for possible nonlinear effects and interactions.

The generalized additive model (GAM) (Hastie and Tibshirani, 1990; Lin and Zhang, 1999; Wood, 2006, 2011; Wood et al., 2015) extends the linear model with tools for modeling nonlinear functional relations between a response variable and one or more predictors. GAMs are especially useful for data where the precise nature of these functional relations is not known. GAMs provide spline-based smoothing functions that take one or more predictors as input and construct wiggly curves or wiggly (hyper)surfaces. Spline smooths are set up such that a proper balance is found between staying faithful to the data and model parsimony. This is accomplished by penalizing smooths for wiggliness.

The effective degrees of freedom (edf) of a smooth, which are used to evaluate the significance of a smooth, reflect the degree of penalization. Penalization may result in all wiggliness being removed from the smooth, resulting in a term

with one effective degree of freedom, in which case the effect of the predictor is linear. Thus, if a predictor has a linear effect, the smooth will simplify to a standard line with a slope parameter. Nonlinear terms in the model are interpreted by plotting the partial effect of the smooth together with confidence intervals. As it is impossible to interpret a non-linear effect from just the model summary, it is essential to always consider the plots of the partial effects. Therefore, plots are used to clarify the size, shape and direction of effects.

The generalized additive mixed model (GAMM) incorporates random-effect factors. When using GAMMs, the modeler has the possibility to replace the combination of random slopes and random intercepts in the linear mixed model, used to model by-participant (or by-item) random variation in regression lines, by wiggly curves. The summary of a GAMM reports both the parametric part of the model (intercept and the betas of the linear terms) and the smooths (wiggly curves and wiggly (hyper)surfaces, as well as random effects). For a brief introduction to GAMMs, see (Baayen et al., 2017a). GAMMs have been used in previous (psycho)linguistic studies, and have been applied to, for example, dialectological data (Wieling et al., 2014) and experimental data (Winter and Wieling, 2016; Baayen et al., 2016b; Van Rij et al., 2016). We used the `mgcv` package (Wood, 2006) for fitting GAMMs to our experimental response variables. For some of the models reported below, the residuals showed thick tails. Here, we dropped the assumption that the errors are normally distributed and instead modeled the scaled residuals as following a t-distribution.

In our analyses, we checked all numeric predictors for non-linearity. Predictors with strictly linear effects can be identified in the model summaries as smooths with only 1 effective degree of freedom (edf). By-subject factor smooths for trial (the rank of a trial in the experiment) were used to model the ebb and flow of attention in the course of the experiment (see Baayen et al., 2017a, for detailed discussion). Smoothing splines were also essential for clarifying the nature of the effects of the NDL predictors. For wiggly curves, we made use of thin plate regression splines, and for wiggly surfaces, we made use of tensor product smooths.

The statistical models reported below are based on exploratory data analysis. From highly correlated predictors, only the one predictor that explained most of the variance was included. Two-way interactions were explored systematically.

4.4 Eye-tracking experiment

Eye-tracking has thus far not been used to study lexical bundles — semantically transparent and compositional multi-word units.[2] Previous eye movement research on multi-word expressions has focused on idioms (Siyanova-Chanturia et al., 2011a; Underwood et al., 2004) and binominal expressions, such as *bride*

[2]With the exception of our study on differences in reading lexical bundles between younger and older adults, see **Chapter 2**.

and groom (Siyanova-Chanturia et al., 2011b). A processing advantage of idioms over literal language was found in the number of fixations participants made, where idioms were fixated on less, and in the total reading time, which was shorter for idiomatic phrases than for matched novel phrases. Siyanova-Chanturia et al. (2011b) presented participants with binominal phrases in their prototypical form, e.g. *bride and groom*, and in their reversed form, *groom and bride*. All phrases were matched on single word frequency, and only differed in phrasal frequency. They found that phrasal frequency significantly affected the number of fixations made, the total reading time, and the first pass reading time, a measure that sums all fixation durations before the first regression is made.

For this study, we focus on the first fixation durations, which reflect the first stage of reading, the first pass reading times, which reflect early processing during reading, and the number of fixations, which reflect the overall difficulty of processing during the whole reading process.

4.4.1 Materials

We randomly selected a set of three-hundred trigrams from the top one percent most frequent trigrams in the Netherlands Dutch part of the OpenSoNaR corpus of contemporary Dutch (Oostdijk et al., 2013). We specifically selected a subset from the most frequent combinations of three words so as to make sure that the stimuli selected were very likely to be stored under any usage-based account (Goldberg, 2003; Bybee, 2010).

The trigrams selected were all semantically transparent combinations of words, so that the meaning of the whole is not idiomatic or opaque, but composed of the meanings of the separate words. These types of multi-word units are often referred to as lexical bundles in the literature (Wray, 2012; Tremblay et al., 2011). Moreover, we did not limit the set of stimuli by choosing only constituents, or combinations of words that can stand alone as utterances. Arnon and Cohen Priva (2013); Tremblay and Baayen (2010); Tremblay et al. (2011) have all shown that phrasal frequency effects appear regardless of whether or not a multi-word unit is a constituent. Nevertheless, we included a predictor in our models specifying if a multi-word unit is a constituent or not, to further test if constituency plays a role in multi-word unit processing.

4.4.2 Design

The experiment started with a practice block of five trials, where each trial was followed by a comprehension question. The rest of the experiment consisted of three blocks, containing 100 trials each. These blocks were separated by short breaks. At random intervals, experimental items were followed by a string of words that was either a grammatical continuation or an incorrect continuation of the trigrams. Participants had to click with a mouse on a 'correct' or 'incorrect' label on the screen and received direct feedback on their choice. One third

of the experimental items was followed by these comprehension questions.

4.4.3 Participants

We recruited thirty-two students from Leiden University (20 female, average age 21.8 years). All participants were native speakers of Dutch and had normal or corrected-to-normal vision. Due to technical issues data from two participants had to be discarded. Participants gave informed written consent prior to participating and they received a monetary reward for their participation.

4.4.4 Procedure

Participants were seated in a sound-proof room. They received verbal instructions about the task, which was reading the trigrams presented on the screen silently, and to answer a set of comprehension questions that were presented at random intervals. The eye movements of their dominant eye were recorded with an Eyelink 1000 eye-tracker (SR Research Ltd). We used a 500 Hz sampling rate and performed eye calibration at the beginning of the experiment, using a 9-point calibration procedure. To minimize head movements, we asked participants to put their head on a head rest. After calibration was achieved, participants received final written instructions on the screen before the experiment started.

At the start of each trial, a fixation point was presented for 500 ms at the left-hand side of the screen, to ensure that they would read from left to right. Trigrams were presented in a black, monospaced font (Consolas, size 22) against a white background for 1,200 ms. One third of the trigrams was followed by a comprehension question, that stayed on the screen until the participant clicked on a box with 'correct' or 'incorrect' with a mouse. Trials were separated by an inter-stimulus interval of 1,000 ms.

4.4.5 Analyses

In order to understand how readers process trigrams, we looked at several eye-tracking measures that reflect different processes over time. To gauge what is happening at the very first moment readers encounter a trigram, we modeled the first fixation durations. Previous research has shown that whole-form frequencies of complex compounds can already influence this early measure (Kuperman et al., 2009; Miwa et al., 2017; Pollatsek et al., 2000). In order to approach normality, we raised the first fixations duration to the power 0.2. The results of the modeling are discussed in Section 4.4.6.

The first fixations durations are not fully representative for early processes in reading of units that consist of several words (Carrol and Conklin, 2015). Therefore, to get a more complete picture of early processing of written trigrams, we also considered the first pass reading times (see Section 4.4.7). First pass reading times represent the duration from the start of the first fixation

until the first regression is made. This measure gives an indication of the processes employed during the initial reading of the trigrams. In order to approach normality, we took the square root values of the first pass reading times.

We also looked at the number of fixations participants made. This measure is thought to reflect processing difficulty. The harder a text is, the more fixations a reader makes (Rayner, 1998). Additionally, it is a measure that provides a summary of the full reading process, giving an indication of what happened during the whole course of reading. To model the fixation counts, we used a generalized linear model with a Poisson link.

4.4.6 First Fixation Durations

The model of the first fixation durations contains significant main effects for the length of the trigram, interactions of the horizontal position of the first fixation (firstFixX) with the age of the participant and the NDL prior (logPrior), and the log frequencies of the third word of the trigrams. There are furthermore random intercepts for items (trigrams), factor smooths of trial number per participant, and by-participant random slopes of the trigram length, fixation position, the frequencies of the single words, and the NDL prior. Only the latter did not reach significance, but was kept in the model as the NDL prior is included in an interaction term. Table 4.1 reports the results.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	3.0788	0.0596	51.6781	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(length)	1.0001	1.0002	79.0230	< 0.0001
te(firstFixX,age)	14.0521	16.0490	6.2623	< 0.0001
te(firstFixX,LogPrior)	3.9835	4.6942	5.6477	0.0009
s(logFreqC)	3.8941	4.5074	4.9069	0.0003
s(trigram)	94.4559	289.0000	0.4866	< 0.0001
s(trial,ptc)	72.3468	268.0000	15.5696	< 0.0001
s(length,ptc)	21.0110	29.0000	6.9429	< 0.0001
s(firstFixX,ptc)	24.5172	28.0000	20.7602	< 0.0001
s(logFreqA,ptc)	2.8022	30.0000	0.1047	< 0.0001
s(logFreqB,ptc)	12.7689	30.0000	1.5642	0.0035
s(logFreqC,ptc)	13.8035	29.0000	1.4906	0.0027
s(LogPriorptc)	0.0004	29.0000	0.0000	0.8092

Table 4.1: Table of the results of the model of first fixation durations.

Figure 4.2 displays the fixed effects of the model. The upper left panel shows how longer trigrams elicit shorter first fixation durations. When a reader encounters a long trigram, it is unlikely that she will be able to process the whole trigram already at the first fixation, and so she will re-fixate as quickly as possible. If the trigram is short, however, then the reader will be able to see most if not all of the trigram from her foveal and parafoveal view (Rayner,

1998), and as a consequence, will not re-fixate quickly.

The top right panel displays the interaction of the first fixation location and the age of the participant. When the first fixation lands near the beginning of the trigram, this fixation tends to be very short, especially so for older participants. However, a similar eye-tracking study with a younger group of participants in their twenties and an older group of participants in their sixties, did not find any age-related effects, despite the much larger differences in age (Lensink et al., submitted). It is not clear if the absence or presence of an age effect is due to false negatives or false positives. It could be the case that due to a larger experience with reading, the older participants in this study were quicker to realize that they need to re-fixate when their first fixation lands near the beginning of the trigram.

If the first fixation landed further into the trigram, however, then the first fixation lasted longer, as there is more information that can be extracted from the signal from that position. For older participants, this effect was even larger. Again, it might be the case that the larger reading experience of older readers makes them better at estimating what the optimal fixation duration is at a certain location, so as to extract as much information as possible.

The bottom left panel shows the interaction of the fixation location with the NDL prior. In this interaction, the further the first fixation landed into the trigram, the shorter this fixation will last. For fixations near the beginning of the trigram, a higher NDL prior will speed up processing, leading to shorter fixations. If the fixations landed near the end of the trigram, the effect flips, and larger NDL prior values correspond to longer fixation durations.

Lastly, the panel on the bottom right shows how the effect of the frequency of the third word has a quadratic shape: First fixation durations tend to get longer only for trigrams where the third word has a log frequency near zero.

It is interesting to see that already at the very first fixation, participants employ top-down information of the full trigram (the NDL prior) and the frequency of the third word. We expected the NDL prior to have a facilitative effect on reading measures, such that higher prior values would correspond to shorter fixation durations. However, when the first fixation lands far enough into the trigram, we see that higher priors correspond to longer fixations. In Section 4.4.9 we will get back to this unexpected finding.

4.4.7 First Pass Reading Times

First pass reading times are the total durations of all reading that happens before readers make a regression. They reflect early processing and are especially useful when considering multiple words at once (Carrol and Conklin, 2015).

The model for the first pass reading times (Table 4.2) contains a significant effect of age, where older readers spend more time on their first passes. The position of the first fixation (firstFixX), trial number, the frequency of the second word, and the NDL trigram activation (LogActTrig) form the significant main effects of the model. Strikingly, the length of the trigram did not reach

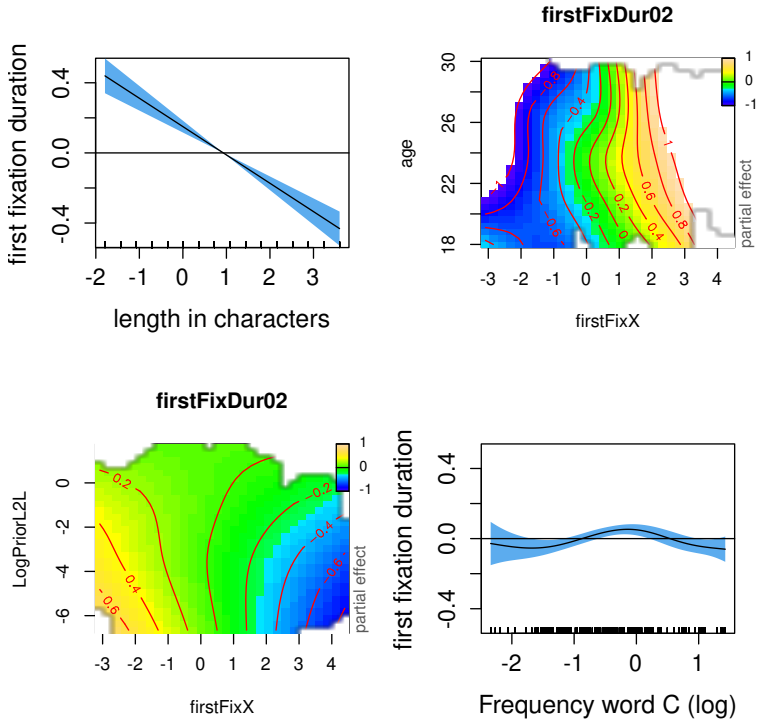

Figure 4.2: Partial effects of the model of the First Fixation Durations. The panel on the top left shows the effect of the lenght of the trigram, the panel on the top right shows the interaction of participant age and the horizontal location of the first fixation. The bottom left panel shows the interaction of the first fixation location and the NDL prior. The panel on the bottom right shows the effect of the third word frequency.

significance and model comparisons showed that it did not have to be included as a main effect in the model. However, there is a significant random slope of length per participant, showing that participants did differ among themselves in how their first pass reading times were influenced by the length of the trigram.

The random effects part of the model contains random intercepts for subjects (ptc) and items (trigram), factor smooths of trial number per participants, and random slopes for the fixation location, the length of the trigrams in characters, the frequencies of the single words, and the NDL trigram activations. Only the latter did not reach significance, showing that there are no significant individual differences between participants in how their first pass reading times are affected by the trigram activations.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	8.5570	4.5979	1.8611	0.0628
age	0.5861	0.2106	2.7838	0.0054
B. smooth terms	edf	Ref.df	F-value	p-value
s(firstFixX)	5.3022	6.4162	64.8170	< 0.0001
s(trial)	1.0003	1.0005	9.7653	0.0018
s(logFreqB)	1.0001	1.0001	10.7538	0.0010
s(LogActTrig)	3.8976	4.4551	2.7768	0.0142
s(ptc)	9.1028	28.0000	0.6302	< 0.0001
s(trigram)	112.1411	279.0000	0.6793	< 0.0001
s(length,ptc)	13.9800	30.0000	1.4194	0.0043
s(firstFixX,ptc)	19.8131	29.0000	3.3395	< 0.0001
s(trial,ptc)	75.6200	268.0000	66.1222	0.0279
s(logFreqA,ptc)	13.0652	30.0000	1.0319	< 0.0001
s(logFreqB,ptc)	13.1648	29.0000	1.1846	0.0045
s(logFreqC,ptc)	8.7518	30.0000	0.5420	0.0600
s(LogActTrig,ptc)	3.2482	29.0000	0.1287	0.2382

Table 4.2: Table of results of the model of first pass reading times.

The partial effects are plotted in Figure 4.3. The first pass reading times tend to get longer over the course of the experiment, which could indicate fatigue (Baayen et al., 2017a). There is a negative direction to the effect of the location of the first fixation on the first pass reading times: When the first fixation landed near the beginning of the trigram, readers spent more time at their first pass than when the first fixation landed near the end of the trigram. This makes sense, as the first pass includes all fixations before the first regression is made — if the first fixation landed near the end of the trigram, then a reader cannot make many forward fixations, so a regression is likely to take place already at the second or third fixation, reducing the time of the first pass.

Higher frequencies of the second word of the trigram correspond to shorter first pass reading times, showing the expected facilitation and shorter reading times for high frequency items (Rayner, 1998). Higher bottom-up activations,

however, correspond to longer first pass reading times. Note that it is mostly the lower values of the NDL activations that have a clear effect on the reading times. As we expected to see facilitative effects of the NDL activations, this is unexpected, and we will further discuss this in Section 4.4.9.

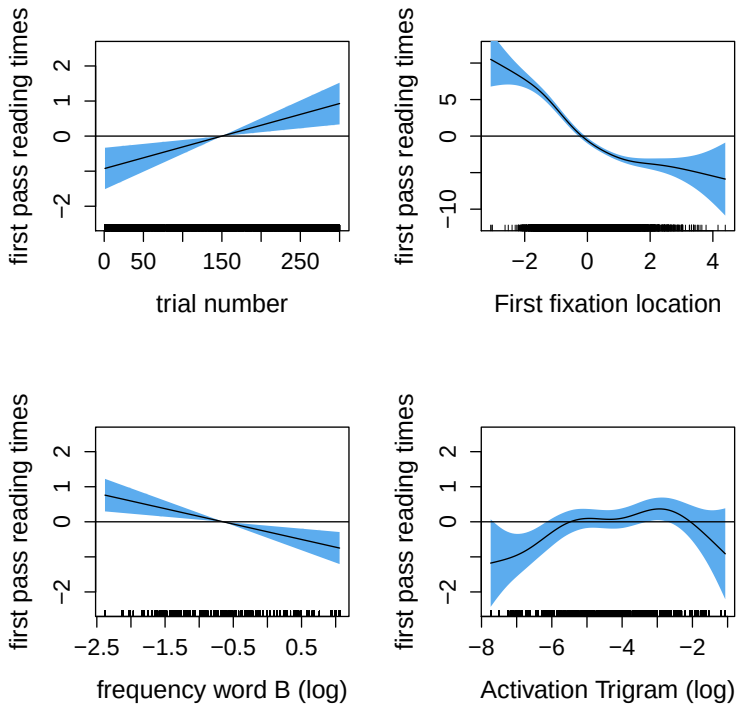

Figure 4.3: Partial effects of the model of the First Pass Reading Times. The top two panels show the effects of the trial number — reflecting the temporal position in the experiment — and the position of the first fixation. The bottom two panels show the effects of the second word frequencies and the trigram activations.

4.4.8 Number of fixations

The number of fixations that participants made on each trigram can tell us something about the overall course of processing. Ease of processing is reflected in a lower number of fixations made (Rayner, 1998).

Table 4.3 shows the results of the model. There are significant main effects for the locations of the first and second fixations, the durations of the first stage of processing — the first pass reading times —, and the frequency of

the first word of the trigram. The effect of the length of the trigram is near significant. There are furthermore significant random intercepts for subjects (ptc) and items (trigram), and non-significant random slopes per participant of the length of the trigram, the locations of the first and second fixation, the first pass reading times, and the frequencies of the first words of the trigrams. Note that none of the NDL measures reached significance, and only the frequency of the first word of the trigram influences the number of fixations made.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1.3214	0.0311	42.5068	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(length)	1.0001	1.0001	3.7318	0.0534
s(firstFixX)	1.1504	1.2823	72.0942	< 0.0001
s(secondFixX)	2.0572	2.6427	53.8001	< 0.0001
s(firstPassRT)	4.5627	5.6345	362.4722	< 0.0001
s(logFreqA)	1.0000	1.0000	6.3252	0.0119
s(ptc)	23.4770	29.0000	127.8936	< 0.0001
s(trigram)	214.2674	264.0000	1217.8949	< 0.0001
s(length,ptc)	0.0000	29.0000	0.0000	0.9997
s(firstFixX,ptc)	0.0001	29.0000	0.0001	0.6687
s(secondFixX,ptc)	1.2191	29.0000	1.3156	0.4117
s(firstPassRT,ptc)	0.0000	29.0000	0.0000	0.8698
s(logFreqA,ptc)	0.0010	29.0000	0.0009	0.4946

Table 4.3: Table of the results of the model of the number of fixations.

In Figure 4.4 the main effects are plotted. The near significant effect of the length of the trigram shows an upward trend, where longer trigrams elicit more fixations. The effects of the horizontal locations of the first and second fixations are each other's opposite: The further the first fixation landed into the trigram, the less fixations overall participants made; the further the second fixation landed into the trigram, the more fixations participants made. This seems to suggest that reading a trigram is optimal when the first fixation lands relatively far into the trigram, and when the second fixation lands relatively near the beginning of the trigram — in other words, when readers make a regression.

How the first stage of processing proceeds, has a large influence on the overall reading process, as shown by the large effect that the duration of the first pass reading time has on the number of fixations made. The longer the first pass lasted, the less fixations readers will need overall. The more time a reader spends at the first stages of processing, the less fixations in total she will need, which is an indication of ease of processing. In other words, it pays off to take more time at the initial stages of processing a written trigram.

The frequency of the first word of the trigram, lastly, has a facilitative effect, such that more frequent first words correlate to less fixations overall.

It is striking that only the first word frequency plays a role in how many fixations readers make, especially since the large majority of the first words of our stimuli are function words. We will come back to this at our discussion of the eye-tracking data in the next section.

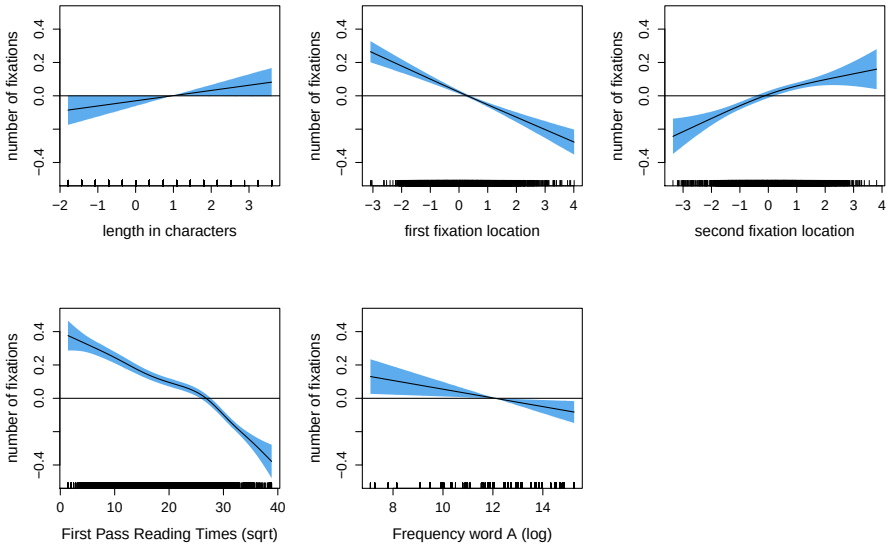

Figure 4.4: Partial effects of the model of the number of fixations. The top panels show the effects of the length of the trigram, and the horizontal locations of the first and second fixation. The bottom panels show the effects of the duration of the first pass reading times, and the frequency of the first word of the trigram.

4.4.9 Discussion eye-tracking data

The eye-tracking data show that the NDL measures provide additional insights over and above more traditional measures of lexical processing, especially for the early stages of reading. The NDL priors and NDL activations explain more of the variance in the data than trigram frequencies, and provide moreover a more nuanced picture of how reading trigrams proceeds. The reader starts with top-down information and continues to process the text using more bottom-up information, where the time spent at the initial stages of reading are predictive of how easy the overall reading process will be.

Already at the durations of the very first fixation, the NDL priors play a role in processing. Recall from Section 4.2 that the priors reflect how active a trigram is when there is no visual input. This could be conceptualized as a

type of resting-state activation (Milin et al., 2017). We expected to see that higher prior values would lead to easier processing and thus to shorter fixation durations. The priors are employed as soon as the first piece of information is perceived by the reader. They have a facilitative effect, i.e. higher priors lead to shorter fixation durations, when the first fixation lands near the beginning of the trigram. However, when the first fixation lands more towards the end of the trigram, then higher priors lead to longer first fixation durations.

This unexpected effect shows that high prior values do no necessarily lead to shorter first fixation durations. This shortening only happens when this first fixation lands near the beginning of the trigram. From this location, the reader is likely to be able to see the full trigram from his foveal and parafoveal view, especially since the parafoveal view of readers of languages that are written from left-to-right is asymmetrical and larger on the right-hand side (Rayner, 1998). This provides the reader with enough visual information to gain facilitative effects from higher prior values. However, when the first fixation lands more toward the end of the trigram, then the reader is not likely to see the full trigram at once, missing useful visual input especially from the beginning of the trigram. From this suboptimal viewing position, perceiving parts of trigrams that moreover have low prior values, will prompt the reader to re-fixate as quickly as possible, as he will not be able to gain much information during that fixation. However, if the prior values are high, then the reader will attempt to process the visual information, and spend a bit more time at the first fixation.

The first pass reading times reflect a further stage in processing, that sums up what happens at the initial stages of processing, before readers make a regression to reread, re-evaluate, or reconsider text that they have read before. Instead of measures reflecting resting-state activations, which could be seen as top-down influences, now the NDL activations start to play a role. These NDL activations reflect bottom-up processes, in this case the bottom-up support from the visual signal to the trigram outcomes. So at the first fixation, readers begin using top-down expectations, and along the way start to use bottom-up input.

We predicted that the NDL activations have a facilitative effect on processing. For the first pass reading times, however, we see that higher activations correspond to longer first pass durations. It seems to be the case that readers prefer to spend more time at the early stages of processing when the visual input provides them with stronger support for a know trigram, in order to try to perceive and process as much information as possible, as early as possible. This explanation fits with the large influence that the durations of the first pass reading time have on the total number of fixations made, which reflects the overall reading process — longer first pass reading times lead to less fixations overall. Previous research has moreover found similar reading strategies for lexical bundles (Lensink et al., submitted). To conclude, there is a clear trade-off of the amount of time spent at the first stages of reading, and the total cost of reading and processing the whole trigram, where a longer first stage corresponds to easier processing overall.

A final remark is in place about the role that the frequencies of the single words play. Even though it is clear that readers use the full trigram from the first fixation onwards, they also make use of the single word frequencies. At the first fixation, there is an effect of the frequency of the third word, at the first pass reading times, there is an effect of the frequency of the second word, and at the total number of fixations, there is an effect of the frequency of the first words. It could be that at the first fixation, participants focus more on the end of the trigram as a way to check if their top-down expectations match reality, and that over the course of processing, they focus more on the middle and beginning of the trigram.

4.5 Production experiment

Moving further into the processes that underlie reading out loud, we now consider the processes giving rise to differences in naming latencies of trigrams and their production durations. Phrasal frequency effects have been well-established for production data (Arnon and Cohen Priva, 2013; Arnon and Priva, 2014; Tremblay and Tucker, 2011). Previous work has largely only looked at English. We extend previous research by replicating these types of experiments with another language, Dutch. We use a word-reading paradigm, where participants are instructed to read Dutch multi-word units out loud from a computer screen. Our prediction is that phrasal frequency will also have a significant effect on production durations of Dutch multi-word units. We moreover explore if, over and above the frequencies, the NDL priors, activations and activation diversities play a role.

4.5.1 Materials

We used the same set of stimuli as used in the eye-tracking experiment (see Section 4.4). We created two new experimental lists, taking care again to ensure that items with phonological or semantic overlap did not precede or follow each other in two consecutive trials.

4.5.2 Design

Two different experimental lists were created, consisting of three blocks of one hundred items, where no trigrams following each other within two trials had any phonological or semantic overlap. The two lists were assigned randomly to the participants. See the online appendix for a full list of the stimuli and the two experimental lists. The three experimental blocks, each consisting of 100 trials, were preceded by a practice block of five trials. All blocks were separated by a short break.

4.5.3 Participants

Thirty students from Leiden University were recruited to participate in the study (21 female, average age 22.0 years). All were native speakers of Dutch. Participants gave their written consent before the start of the experiment and received a monetary reward for their participation.

4.5.4 Procedure

Before the start of the experiment, participants were given written information about the experiment and they gave their written consent. Participants were asked to read out loud the words on the screen as fast and as accurately as possible. First a fixation cross was presented in the middle of the screen (font: Arial, size: 18) for 500 ms, followed by a 100 ms blank screen. Then a trigram was presented (font: Arial, size: 18) for 1,200 ms. All letters were printed in black against a white screen. Each trial was separated by an inter-stimulus interval of 1,000 ms. A microphone recorded the speech of each participant.

4.5.5 Analyses

In order to gain more insight in the processes active during the production of trigrams, we considered the onset latencies that mark the beginning of the utterances, and the total durations of those utterances. Both dependent measures were log-transformed to approach normality.

4.5.6 Production onset latencies

When reading a trigram out loud, it makes a difference if this trigram is a constituent or not, as shown by the significant effect that constituency has on the onset latencies (see Table 4.4). A participant that has to read out loud a trigram that is a constituent is a bit slower in starting to speak than a participant that has to read out loud a trigram that is not a constituent.

Next to an effect of constituency, the model contains a near significant effect of trial number and a near-significant interaction of the NDL activations and the NDL activation diversities. There are moreover significant main effects of the length of the trigram and the single word frequencies. The model includes random intercepts for items (trigrams), factor smooths of trial per participant, and random slopes for the NDL activations, the NDL activation diversities, and the single word frequencies, which all reached significance.

Figure 4.5 shows how speakers got a bit faster over the course of the experiment, how longer trigrams take longer to read out loud, what the interaction between the NDL activations and the NDL activation diversities looks like, and how higher frequencies of the single words speed up the onset latencies.

The plot in the upper right corner, displaying the interaction of the two NDL measures, shows that larger NDL activations tend to speed up naming. The

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-1.9596	0.0584	-33.5792	< 0.0001
constituentY	-0.0631	0.0195	-3.2410	0.0012
B. smooth terms	edf	Ref.df	F-value	p-value
s(length)	3.1420	3.2463	9.3388	< 0.0001
s(trial)	2.7254	2.9455	2.3240	0.0621
te(LogActTrig,logActDiv)	5.6942	5.8812	2.0514	0.0681
s(logFreqA)	2.1530	2.2256	16.2560	< 0.0001
s(logFreqB)	1.0002	1.0002	7.6802	0.0056
s(logFreqC)	3.4033	3.5162	5.8700	0.0004
s(trigram)	212.9775	258.0000	5.6714	< 0.0001
s(length,Ptc)	15.1292	29.0000	1.3553	0.0001
s(trial,Ptc)	177.0317	269.0000	4426.6984	< 0.0001
s(LogActTrig,ptc)	9.0629	29.0000	0.4712	0.0497
s(logActDiv,ptc)	6.0253	29.0000	0.2730	< 0.0001
s(logFreqA,ptc)	14.6149	29.0000	1.2567	0.0013
s(logFreqB,ptc)	20.4309	29.0000	4.0641	< 0.0001
s(logFreqC,ptc)	13.9648	29.0000	1.5245	0.0003

Table 4.4: Table of results of the model of the production onset latencies.

better the bottom-up support is, the better participants can prepare themselves for articulation, and the faster they will start speaking. This facilitative effect of NDL activations is strongest for trigrams with high activation diversities.

The NDL activation diversity is a measure that conceptually resembles measures of neighborhood density. The larger the diversity, the larger the number of other outcomes that are also supported by the cues in the input. This leads to more difficulty in processing, which in turn could lead to delayed onsets and larger durations. However, this inhibitive effect of activation diversities is only seen for trigrams with very low activation values. For trigrams with moderate or higher activation values, higher diversity values lead to faster naming. So when the visual input supports a lot of different possible trigrams, and when this is accompanied by an moderate to large bottom-up support for the intended trigram too, then the participant will start speaking faster. We will get back to this result in Section 4.5.8.

4.5.7 Production durations

Whether or not a trigram is a constituent has no influence on the production durations of a trigram. There are significant main effects of the length of the trigram, trial number, the frequencies of the first and second word, and the trigram frequencies in our model. NDL measures did not reach significance as main effects in the model, but do play a role in the random effects structure. This means that individual participants differ significantly in how their production durations are influenced by NDL activations and NDL activation diversities, but that there was no overall effect of these measures. See Table 4.5 for an overview.

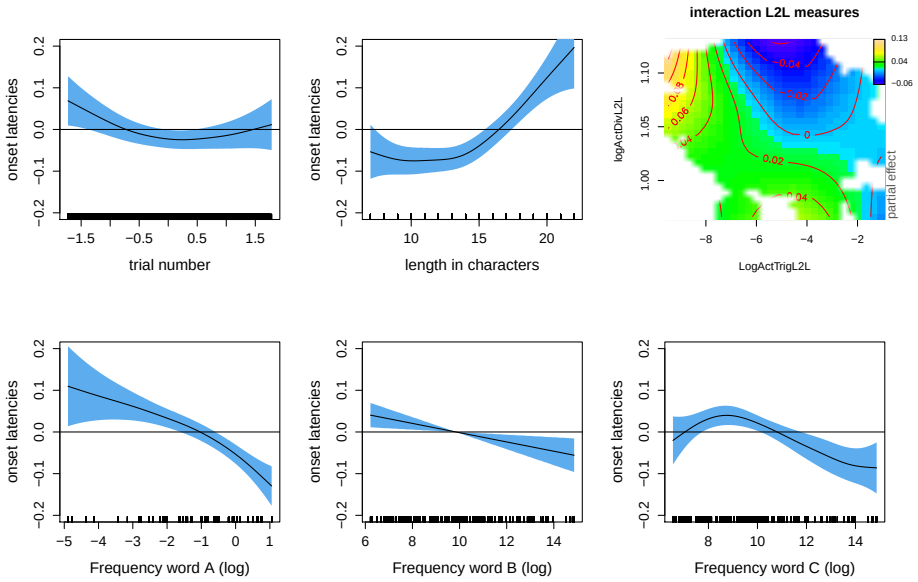

Figure 4.5: Partial effects of the model of the Onset Latencies of the production data. The first two top panels show the effects of trial number and length of the trigram. The panel at the top right shows the interaction of the NDL trigram activation (logActTrig) and the NDL trigram diversity (logActDiv). The bottom three panels show the effects of the single word frequencies.

The model furthermore includes random intercepts for items (trigrams), factor smooths of trial number per participant, and random slopes per participant of the length of the trigram, the NDL activations, the NDL activation diversities, the single word frequencies, and the frequencies of the full trigram.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.7543	0.0256	264.2721	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(length)	1.0003	1.0003	518.5208	< 0.0001
s(trial)	1.0000	1.0000	4.2511	0.0393
s(logFreqA)	2.6453	2.6638	2.9704	0.0318
s(logFreqB)	1.0001	1.0001	5.6982	0.0170
s(logFreqABC)	1.0001	1.0001	12.9681	0.0003
s(trigram)	251.7944	262.0000	32.5657	< 0.0001
s(length,ptc)	23.6631	29.0000	7.6113	< 0.0001
s(trial,ptc)	186.9775	269.0000	20506.0712	< 0.0001
s(LogActTrig,ptc)	11.1132	30.0000	0.7370	0.0190
s(logActDiv,ptc)	5.7714	30.0000	0.2491	< 0.0001
s(logFreqA,ptc)	11.3527	29.0000	0.9122	0.0141
s(logFreqB,ptc)	22.4854	29.0000	9.9231	< 0.0001
s(logFreqC,ptc)	19.7342	30.0000	5.7716	< 0.0001
s(logFreqABC,ptc)	8.5092	29.0000	0.8089	0.0681

Table 4.5: Table of results of the model of the production durations.

Figure 4.6 shows that production durations get slightly shorter over the course of the experiment, that longer trigrams take longer to pronounce, and the effects of the frequencies of the first two words and the trigram itself. The frequency of the first word has a quadratic shape, with high frequency first words slowing down production durations, which is unexpected. However, the effect is quite small, and might not be robust. The effect of the frequency of the second word goes in the expected direction, with higher frequency second words leading to shorter overall production durations. Lastly, the frequency of the trigram also has a facilitative effect on production durations: The higher the frequency of the trigram, the less time participants need to produce the whole trigram.

4.5.8 Discussion production data

This section seeks to study the processes involved in lexical access of trigrams when people are speaking, and to see to what extent NDL measures could add any new insights over and above traditional measures of lexical processing such as the frequency of an item or its length in characters. The NDL measures play a role in how fast people start to speak, but we did not find any main effects of NDL measures in the production durations.

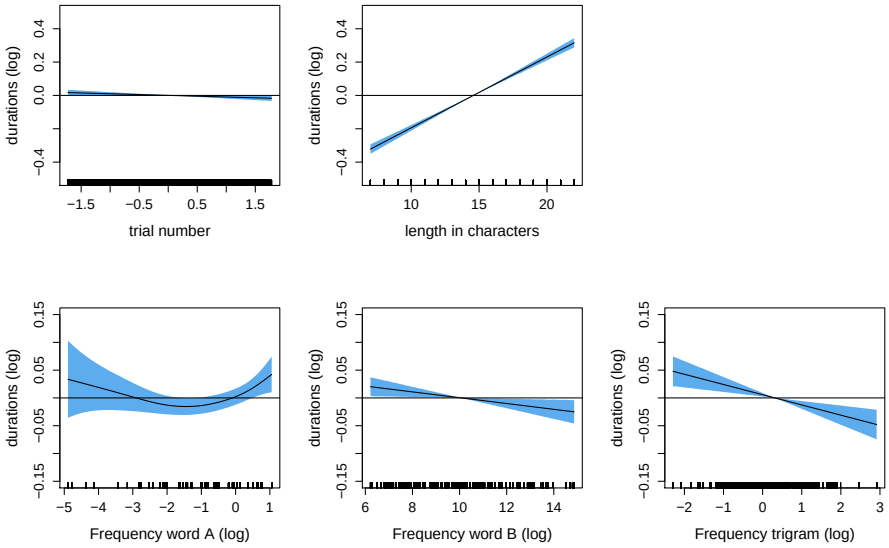

Figure 4.6: Partial effects of the model of the production durations. The top two panels show the effects of trial number and the length of the trigram. The bottom panels show the effects of the first word frequencies, the second word frequencies, and the trigram frequencies.

Arnon and Cohen Priva (2013) found robust trigram frequency effects in their study, irregardless of the constituency of those trigrams. To see if the same applies to Dutch trigrams, we also considered the constituency of the trigram. We found that onset latencies are delayed for constituents, but did not find any effect of constituency on the production durations. Participants are quicker in starting to speak when reading out loud non-constituents. It could be the case that constituents evoke more semantic and pragmatic associations, slowing down the speaker. It could also be the case that constituents prompt the speaker more to use a certain intonation contour, whereas non-constituents can be pronounced with a more monotone intonation. The latter might require less planning and speakers will therefore be quicker to start speaking.

When considering the model of the onset latencies, it appears that all single word frequencies influence how fast people start to speak. This suggest that before speaking, all single words have been recognized and are employed in preparing the utterance. There are however no effects of the full trigram frequencies or the NDL prior on onset latencies, which is unexpected given that we found trigram effects already at the first fixation, and previous work has shown early influences of whole-form compound frequencies (Kuperman et al., 2009; Miwa et al., 2017; Pollatsek et al., 2000).

However, there are trigram effects at play, but these effects are different from traditional frequency measures. There is a small interaction of the NDL activations and the NDL activation diversities, which index the total bottom-up support for the target trigram, and the number of other outcomes that are also supported by the visual input, respectively. The interaction between the NDL activations and NDL activation diversities shows an inhibitive effect of activation diversities for trigrams with very low activation values. So when the visual input only weakly supports the target trigram, and when there is a large uncertainty about the identity of the trigram, participants are slowed down. However, when the visual input provides moderate to strong support for the target trigram, then larger activation diversities lead to faster onset latencies. This could indicate that a larger activation of similar candidates aids in processing, by means of spreading activation from the non-target trigrams to the target trigram, promoting the articulatory processes needed for its production.

Tremblay and Tucker (2011) conducted a production study where participants had to produce frequent four-word sequences. The authors found that the onset latencies in their data were mostly influenced by log probability of occurrence, which they interpreted as indicating a competition of the target multi-word unit with its family members. As the NDL activation diversities are conceptually similar to measures of neighborhood densities, this fits well with our finding that the NDL activation diversities influence the onset latencies in our data. However, Tremblay and Tucker (2011) found that trigrams were the most important predictor for the onset latencies, whereas single words formed the most important predictors for the production durations. We did find trigram frequency effects in the production durations, and only a small interaction effect of trigram activation and diversities measures. That said, Tremblay and

Tucker (2011) also took into account the effects of bigram (AB and BC) and skipgram (AC) frequencies, which we did not. This could explain the difference between the results. For future studies, it will be interesting to also look at the effects of bigrams and skipgrams using a discriminative approach.

4.6 General discussion

We started off by proposing that multi-word units are a feasible theoretical construct. If we however do assume that multi-word units are units of processing, we can ask the questions why these units exist, what they are, and how these units can be discriminated from each other in lexical access.

As to the usefulness of multi-word units as a theoretical construct for gauging lexical processing, we pointed out that lexical storage is extremely rich. Moreover, most multi-word units used in experiments are actually semantic units of their own that encode more than just the sum of their parts. They encode time markers such 'on the day', discourse markers such as 'I think that', and affordance relations such as 'on the table'. As for the question pertaining to the lexical access of multi-word units, we took a discriminative learning perspective to explore to what extent these multi-word units can be discriminated from orthographic input. The computational implementation of this learning perspective, NDL, incorporated multi-word unit lexomes as outcomes. Allowing for multi-word unit lexomes assumes that there is no principled difference between these units and single words, which were posited as outcome units in previous NDL models (Baayen et al., 2011). We predicted that if multi-word units are indeed units, then measures predicting a phrasal frequency effect should also arise in a discrimination model for lexical access to these units. Indeed, we found that the priors taken from the network are very similar and show a high correlation ($r = 0.96$) to trigram frequency values taken from a corpus. Furthermore, the NDL network offers us measures that quantify the amount of activation a trigram receives from the orthographic input (activations) and the uncertainty about the identity of a trigram (activation diversities). We have shown that in silent reading and reading out loud of multi-word units, these measures add additional insights over and above frequency values. These results testify to the plausibility and usefulness of a discriminative approach.

Moreover, by including NDL and frequency measures pertaining both to single words and trigrams, and the location of the fixations as predictors in our model, we also tackled the methodological issue of how to use eye-tracking to study units that are simultaneously compositional strings and whole units, each of which has their own set of factors influencing reading behavior (Carrol and Conklin, 2015).

The question remains why we only found clear effects of the NDL measures in the eye-tracking data, a small interaction effect in the onset latencies of the production data, and no main effects of NDL measures in the production durations. Recall that we aimed to study how lexical access of multi-word units

proceeds. NDL networks provide us with new measures of lexical access, i.e. priors, activations, and activation diversities. Lexical access occurs during the first stage of reading, but will become less active and important over time — explaining why NDL measures of lexical access do not play a role in the model of the total number of fixations made. The same applies to the production data, where at the time of the onset of articulation, speakers are still influenced by measures of lexical access, whereas further down the production process, lexical access has already taken place and its measures do not play any significant roles anymore for most speakers in the total production durations.

4.6.1 Lexical access of multi-word units

Our data show that lexical access to multi-word units proceeds from top-down processes as indexed by the NDL priors, to bottom-up processes where the support for a certain multi-word unit from the visual input goes hand in hand with processes of lexical neighborhoods as indexed by the activation diversities. When reading a trigram, readers are at first influenced by the top-down NDL priors, and then by the NDL activations. It pays off to spend more time at the first pass, by taking the time to let bottom-up visual input inform processing — as indexed by the positive slope of the effect of the NDL activations on the first pass reading times.

When wanting to read out loud a written trigram, speakers will go through at least the first stages of reading before starting to speak. When they are ready to start articulating, it is the frequencies of the single words that speed them up, and trigram measures in the form of bottom-up support and the activation measures. Enhanced bottom-up support speeds up processing, and granted that this bottom-up support is high enough, the co-activation of similar items also aids in processing. The fact that higher trigram frequencies lead to overall shorter production durations, shows that more frequent forms tend to get reduced more in production (Bybee, 2010).

One thing to note is that frequency values outperformed the NDL priors in our production data, whereas these measures produced very similar models and are highly correlated ($r = 0.96$). The reason for this better performance of the frequency values is that NDL priors only capture the form-driven discrimination, whereas frequencies capture more than that. The frequency measures capture two aspects of lexical processing, one relating to the "prior availability" — which is also captured by the NDL priors — and the other relating to higher-order lexical knowledge such as how often things happen in the world, and how these things cluster in the world. This additional layer seems to be more important during production than during reading, where the NDL priors outperformed the phrasal frequency predictors.

For future studies, it will also be worthwhile to see the extent to which trigrams with no clear functional pattern — in contrast to the time markers, discourse markers, and affordance relations mentioned above — can also be implemented as multi-unit words, or single lexomes, in an NDL network. In

this study, we used constituency as a proxy for semantic unity and found that constituency only affected the onset latencies of the production data. Previous research has moreover also indicated that phrasal frequency effects arise irregardless of the constituency of a multi-word unit (Arnon and Cohen Priva, 2013). For future studies, it will be insightful to clearly define what would constitute a 'semantic unit' and contrast semantic with non-semantic units. If the function of the trigrams drives their coherent, single form, then it is expected that trigrams that lack such a coherent function are not processed as chunks. It is also worthwhile to use more sophisticated features than single words as cues, such as the frequency band summary features used by Arnold et al. (2017). In modeling reading, we could implement cues that consist of sub-graphemic orthographic features, which are known to play a role in reading (Dehaene, 2009; Linke et al., 2017).

Overall, this study has shown that incorporating multi-word units as single-unit outcomes in an NDL model works well in predicting empirical data. Moreover, it leads to more insights into the nature and processing of multi-word units. Both single words and the full trigram, their frequencies, priors, bottom-up activations and the activation diversities play a role in lexical access of trigrams. The fact that the NDL approach is successful, hints at the possibility that single words, idioms, and multi-word units are essentially the same type of entity cognitively. NDL theory proposes that linguistic categories, such as morphemes, words, and phrases, are all emergent from a system that simply discriminates between linguistic encodings of relevant pieces of experiences (Ramscar, 2013; Baayen et al., 2017b; Ramscar and Port, 2015; Baayen et al., 2016a). Sometimes these experiences are encoded as a morpheme, sometimes as a single word, an idiom, a multi-word unit or even as a whole phrase - and all are units that we need to keep apart. Discrimination measures can enrich our understanding of the processing of all parts of language.

Acknowledgements The authors would like to thank Maxime Tulling for her invaluable help with the experimental set-up and the collection of the experimental data, Martijn Wieling with his help on analyses of earlier versions of this paper, and Kate Bellamy for her comments on the text of previous versions.

CHAPTER 5

Conclusion

Essentially, all models are wrong, but some are useful

George E. P. Box

In this dissertation I investigated the on-line processing of lexical bundles, and did so by reporting on reading and production experiments, statistically modeling this experimental data, and using a computational model of lexical access. The results presented add novel insights to the existing literature on lexical bundle processing, where the main extensions on those previous findings are a) the focus on advanced statistical models that bring to light the subtle intricacies of lexical bundle processing; b) the first data on how older adults process lexical bundles; c) a more in-depth analysis of the time-course of processing spoken lexical bundles; and d) by explicitly modeling lexical access of units larger than a word in a computational model, this dissertation has proposed a way in which lexical access to written lexical bundles (both when reading silently and when reading aloud) might proceed, and has thereby also made claims on the status of lexical bundles in the lexicon.

Overall, this dissertation has shown that, regardless of modality (reading, speaking, and listening), there is a clear frequency effect of units larger than a single word in Dutch. This concurs with the claims made by usage-based models of language, that state that our usage of language shapes the way language is represented in the brain. From this claim, it follows that frequently used combinations of words become chunked over time and might eventually become units in processing, similar to single words. Indeed, the phrasal frequency effects, the similar syntactic structure of the majority of lexical bundles, the

similar time course and processes involved in processing single words and spoken lexical bundles, and the similar processes of lexical access to single words and the spoken and written lexical bundles investigated in this dissertation, all provide evidence in favor of regarding frequent lexical bundles as units similar to single words.

This notwithstanding, this dissertation has also shown that lexical bundles still retain their internal structure, as the frequencies and other features of their constituent single words and bigrams also play a role in processing, next to their trigram frequencies. In other words, even though lexical bundles are processed as wholes, the language system also analyses their internal structure and takes into account their constituent parts in parallel.

5.1 Reading

Chapter 2 investigated to what extent language experience influences the way lexical bundles are read, by testing two groups of participants: People in their twenties and people in their sixties. The main research question asked was **How do adults read lexical bundles, and are there differences in reading behavior between younger and older adults?**. Assuming a usage-based view on language representations, where usage is believed to shape the way language is represented, it is expected that lexical bundles are represented differently in younger and older adults, given that the latter group has a larger experience using lexical bundles. This in turn is expected to manifest itself in differences in reading behavior of lexical bundles.

The data reported in **Chapter 2** did not show any age-related differences in how lexical bundles are read. This suggests that additional language experience has no measurable consequences for how lexical bundles are read, which does not concur with predictions from a usage-based perspective. In a usage-based approach, it is assumed that language experience over time changes the way language is represented in the brain, which in turn is expected to result in different processing strategies in younger and older adults, which might be measured in an experimental study. So at least for the time being, this claim from the usage-based approach, has not been confirmed. It is, of course, also possible that differences between younger and older adults do exist, but are so subtle that they are only measurable using a larger data set, different experimental techniques such as EEG, or only become manifested when people listen to or produce lexical bundles. Moreover, it should be taken into account that older adults have larger lexicons, which means that they have a larger search space to go through, which in turn will slow down processing overall. Even when older adults might be faster at processing lexical bundles, their longer search through the lexicon might flatten out any of the processing benefits. In other words, we could be facing a ceiling effect, the mechanics of which are still unknown to us. Future studies could help in further investigating whether language experience changes the way lexical bundles are processed.

The data did show effects of trigram frequencies, already at the first fixation durations. Interestingly, these trigram frequencies show an Inverted Frequency Effect, where higher frequency trigrams correlate with longer looking times. These longer early fixation durations in turn correlate with fewer fixations made overall, suggesting that longer early fixations are part of a reading strategy where readers spend more time on their fixations when an item is easy to process, and will spend less time and fewer fixations overall, whereas readers will spend less time at early fixations when an item is difficult to process, quickly re-fixating to get more information, and spending more time and fixations on reading the trigram. As such, longer early fixation durations are indicators of ease of processing, and ease of processing can only be gauged when looking at either later measures such as the number of fixations made, or by considering the whole process from beginning to end.

5.2 Listening

Chapter 3 sought to study how comprehension of spoken lexical bundles proceeds, a process that has not been studied before. Research questions asked were **Is there a difference in electrophysiological brain responses when listening to frequent lexical bundles and infrequent matched controls?**, **Which factors influence the electrophysiological brain response when listening to lexical bundles?**, and **What is the time course of processing of auditorily presented lexical bundles?**.

Two sets of stimuli were created, a list of frequent lexical bundles, and a list of their matched controls. The matched controls were made by replacing the last word of the lexical bundles by a word that is equally frequent, but that forms the end of a less frequent phrase. For example, the lexical bundle *een belangrijke rol* ('an important role') formed the basis of the matched control *een belangrijke dag* ('an important day'), where all single word frequencies were equal, but the second bigram and trigram frequencies differed. Participants listened to recordings of the trigrams read out loud and completed comprehension questions, while an EEG machine collected electrophysiological data.

The ERPs were time-locked to the last syllable of the second word, to capture the moment in time where the lexical bundles started to diverge in terms of pronunciation from their matched controls. The ERPs collected show a sustained negativity, with a clear and widely distributed difference in amplitudes between the conditions. Lexical bundles show less negative amplitudes overall, and start to diverge from the control items at an early point in time. Using a conditional inference random forest analysis, the chapter explores the different roles that a diverse set of predictors has on ERP amplitudes, and how the signal evolves over time.

A result from the random forest model shows three stages in processing. The first stage shows signs of processes of top-down predictions and bottom-up processes initiating the first stage of lexical access. This stage is characterized

by more positive amplitudes for more frequent forms. The second stage involves competition between similar word, bigram and trigram candidates, and the inhibition of similar forms. At this stage, higher frequency first bigrams in lexical bundles correlate with more positive amplitudes, whereas more frequent second bigrams seem to elicit competitory effects and thus more negative amplitudes. The third stage consists of processes of lexical integration, where ease of integration is indexed by a reduced P600 in the form of more negative amplitudes.

Chapter 3 has come up with proposals on how auditory lexical bundle processing proceeds, proposing that top-down expectations take place concurrent with bottom-up signals, and that both single word and bigram frequencies already play a role at the first stage of processing. These parallel influences suggests listeners make use of an interactive comprehension process where different types of information on lexical bundles are employed simultaneously. Moreover, the stages of processing are similar to those of single word processing, suggesting that single words and lexical bundles are quite similar in nature.

5.3 Reading and speaking

Chapter 4 considered new data of both a reading study and a production study of frequent Dutch lexical bundles, and explored to what extent measures from a discriminative learning model of lexical access could add further insights. We know that lexical access to single words involves, among other factors, the frequency of the word, its length, and the properties of its lexical neighbors. Previous research on lexical bundle processing has considered frequencies and length, but did not consider neighborhood densities. The computational model used in this chapter, Naive Discriminative Learning (NDL), provides a proxy for lexical neighborhood effects in the form of an 'activation diversity' measure, which indeed provided additional insights.

The chapter aims to answer the research questions **How does lexical access to lexical bundles proceed?**, **What is their status in the lexicon?**, and **Are there other factors over and above traditional frequency measures that play a role in reading out loud frequent lexical bundles?**.

It is shown that the measures extracted from a discriminative model proved better predictors of the reading measures than traditional frequency measures: The NDL prior and the NDL activations, which represent top-down and bottom-up processes, explain more of the variance in the data than frequency counts. Note that traditional frequency measures are not able to tease apart bottom-up and top-down processes. This makes an NDL approach more insightful as it is more explicit on when information from the written text itself is playing a role, and when top-down information is employed.

The reading study from **Chapter 2** has shown that properties of the whole trigram are already playing a role at the very first fixation durations. Readers are very quick to recognize that they are reading a lexical bundle, and they

are able to access properties of the lexical bundle from an early point onwards. In **Chapter 4**, this process is further teased apart as the data show that readers are at first mostly influenced by top-down expectations. Given that they had to read through a list containing only trigrams, it is not surprising that the participants are primed to expect trigrams, and that they employ top-down processes during the experiment. Initial lexical access of those trigrams is moreover not only determined by top-down expectations, but also by the landing position of the eye on the trigram.

Furthermore, the time spent at the initial stages of reading are predictive of how easy the overall reading process will be. Similar to the results of **Chapter 2**, readers tend to spend more time at the initial stages of reading when an item is easy to process, and will spend less time overall. Although this result has been replicated throughout studies reported on in this dissertation, and has also been found in an eye-tracking study of Japanese lexical bundles (Lensink et al., in preparation), it has not been recorded in the literature before.

After the initial stages, readers start to pay more attention to bottom-up input, as seen in the larger influence of NDL activations. They also shift their attention from the last word of the trigram to the single words at the beginning of the trigram. It seems to be the case that readers first check if their expectations match reality by going over the input on the right-hand side of their foveal vision, and after that focus more on the middle and beginning of the trigram. This seems to go in the opposite direction of the way listeners process spoken trigrams (**Chapter 3)**, who, upon hearing the last word of a lexical bundle, first further process and integrate the beginning of the bundle, before focusing on the last parts. Written lexical bundles are presented at once and as such can be perceived and processed in any given order, whereas sound can only be perceived and thus processed unidirectionally, from the beginning of the lexical bundle to the end.

Note that a small effect of age was found in these data, as opposed to the data discussed in **Chapter 2**, where no age effects were found, even though the age differences between the participants of the study discussed in **Chapter 4** are much smaller than the age differences between the participants of the study discussed in **Chapter 2**. It is not clear yet if the age effects found in **Chapter 4** are true effects, and if the absence of any age effects in **Chapter 2** is due to false negatives. This would mean that the usage-based approach makes correct predictions, and that the absence of an effect in **Chapter 2** is due to type II errors. More research is needed, preferably using different experimental techniques and experiments focusing on speaking and listening.

The production study showed how single word frequencies, NDL bottom-up activations, and an NDL measure of neighborhood density influence onset latencies, whereas the total duration of reading out loud a lexical bundle is mostly determined by the trigram frequencies and to a lesser extent by the frequencies of the first two single words.

5.4 Overall conclusions

Lexical bundle processing proceeds in a similar way as single word processing, but with additional lexical factors, i.e. the properties of trigrams and bigrams, and lexical neighborhood effects based on similar lexical bundles. The way lexical bundles are processed differs between written, auditory, and spoken stimuli, but all three include bottom-up and top-down processes, and influences from smaller parts. The ERP data from **Chapter 3** have moreover shown that lexical access to lexical bundles involves similar stages as lexical access of single words, where after an initial competition among similar forms and an inhibition of non-target lexical bundles, the target lexical bundle is selected for further lexical integrative processes as reflected in the ERP amplitudes.

Besides exploring how lexical access to lexical bundles proceeds and how lexical bundles are processed, **Chapter 4** also discusses why certain transparent combinations of words would and could exist in the mental lexicon. It seems likely that these combinations are not just very frequent by chance alone: the majority of items used as stimuli in this dissertation seem to encode relevant experiences in the world that form either discourse markers such as *I think that*, affordance relations such as *on the table*, and complex time or space markers such as *on the day* and *in the middle of* — items that happen to be expressed as multiple words in languages such as English and Dutch, but that can be encoded as single words in other (morphologically rich) languages.

In **Chapter 2** it was furthermore shown that most stimulus items tend to have very similar structures, with function words forming the first word of the trigram in over 90 percent of the time — this could point to the possibility of a link between language structure, frequency, and semantic unity. Moreover, our conventions to place spaces between certain combinations of sounds are to some extent arbitrary and do not necessarily reflect any grammatical (or even phonetic) reality. As long as we linguists cannot agree on any definition of what exactly constitutes a word, only considering where orthographic conventions have agreed to place white spaces is nowhere near any satisfactory account of what should be considered as a semantic unit that we like to call a 'word'. Concluding, this thesis has provided experimental evidence that units of form and meaning are not necessarily single words or opaque idioms, but could also consist of transparent, frequent combinations of words. By considering certain combinations of words as units of form and meaning, equal to single words, we gain a more realistic view on what the building blocks of language are.

Overall, the eye-tracking data from **Chapter 2** and **Chapter 4**, and the ERP data from from **Chapter 3**, have shown that lexical bundles play a role at several stages in processing, in both production and comprehension, and can be found in eye-tracking data, ERP data, and production data. At first mostly top-down prior expectations play a role in processing, after which the bottom-up input is employed, similar trigrams that have been activated either aid in processing or need to be inhibited, while information from single words, bigram,

and the whole trigram are combined and integrated. This provides additional evidence for a model of language processing where different units are processed in parallel, including units larger than a word, and without distinction between syntactic and semantic processes.

5.5 Useful models - an outlook

A large focus of this thesis is statistical and computational modeling. Therefore, some closing remarks on using statistical modeling as a way to understand the world around us are in order.

We all know the famous adage of statistician George E.P. Box that all models are wrong, but some are useful (see e.g. Box and Draper, 1987). But what constitutes a 'useful' model?

Breiman et al. (2001) distinguishes two different approaches to statistical modeling: One he refers to as the 'Data Modeling Culture', which includes most academic research, and the other one the 'Algorithmic Modeling Culture', which includes most work done in industry. In the Data Modeling community, the focus is on trying to discover the underlying mechanisms that produce the data measured, as a way to better understand the phenomenon at hand. A useful model is understood as a model whose inner workings can be dissected, described, and interpreted. Most importantly, it is often assumed that by using machine learning this way, one can arrive at an approximation of the underlying true mechanisms that cause a certain phenomenon.

This thesis is following this tradition to a large extent by using statistical models that are amenable to such an interpretation: Regression models are transparent in that they show which predictors are more heavily weighted to model the data measured. By visualizing the functional relationship of those predictors with the outcome variables, as done throughout this thesis, it becomes clear how every single predictor adds to producing the phenomenon studied. Making use of a two-layer neural network whose inner workings are relatively easy to capture (the NDL model of **Chapter 4**), is also an example of using a model for its interpretability. The assumption made by many is that using machine learning this way, we will start to understand the true nature of linguistic processing better.

A pitfall of this approach, however, is that it is limited by the imagination of the researcher: As the researcher has to define which might be the relevant factors to input to a machine learning model, it is quite possible that important, unexpected, factors are not included, considered, and discovered, leading to spurious correlations between the factors that the researcher has selected and the data measured. It is quite worrying, at the very least, that there often exist multiple models that fit the data equally well, but that give very different pictures of which predictors are important, and what the relationship between those predictors and the outcome variable are (also known as the Rashomon Effect, see Breiman et al., 2001).

On the other hand, there exists the Algorithmic Modeling Culture. It includes most work done in industry, and includes models such as deep neural networks that are often perceived of as 'black boxes'. Although quite a lot of steps are being taken in the direction of more transparent, 'explainable' models (Samek et al., 2017), it is at the very least quite hard to understand all the underlying rules and heuristics that emerge from the different hidden layers of a deep neural network.

In this Algorithmic Modeling culture, a useful model is not a model that is interpretable and thus explainable, but a model that is as accurate as possible in making predictions, explicitly so without having to approach the way the data has been truly generated in nature. In other words, the focus is not on approaching the truth, but creating a model that works, regardless of the way in which this is achieved. In the last couple of years, these types of models are starting to exceed human performance on tasks such as image classification (e.g. in medical screening) and natural language processing.

Breiman et al. (2001) argues that a model that is better at predicting new, unseen data, is more likely to approach the truth than a model that is as simple and interpretable as possible, even though the model that is better at predicting is less parsimonuous and it is too complex is to completely understand all its inner workings with current tools. We cannot be sure that the mechanisms proposed by a more complex, algorithmic model approach human brain functions, but it is worthwhile to consider the possibility that we might learn new insights from them. "The evolution of science", Breiman argues, "is from simplex to complex" (p. 229 Breiman et al., 2001), and he mentions the developments in the field of physics, where one has moved from Newton's equations to the more complex equations of general relativity, and the emergence of the extraordinarily difficult to interpret equations of quantum mechanics. Despite the complexity of these models, physicists consider them as the current best models of the physical world, and try their best to gain as most knowledge as possible from them.

We live in exiting times, where both computing power and machine learning algorithms are constantly improving, and where the potential to gather and use bigger and more complex data is growing. It is crucial to understand the shortcomings and possibilities of machine learning. Although we should never abandon linguistic theory to guide our questions and interpretations, neither should we shy away from using advanced machine learning algorithms to give us new, unexpected insights. There is still so much we do not understand about one of the most complex behaviors of human beings, language. Let machines assist us in understanding it just a little bit better.

Appendix A

trigram	translation	frequency ABC
aan de man	to the man	15373
aan de universiteit	at (the) university	4073
aan de vooravond	on the eve	7798
aan het begin	at the beginning	89579
aan het eind	at the end	150873
aan het werk	working	1285050
aan te pakken	to deal with	39314
aan te passen	to adapt	41295
achter het raam	behind the window	9642
begin dit jaar	at the beginning of the year	13635
bij grote bedrijven	at big companies	969
bij hem thuis	at his house	10643
bij hun moeder	at their mother's	1754
daar gaan we	there we go	241060
dat blijkt uit	that appears from	1461
dat is altijd	that is always	67085
dat is jammer	that is a pity	42773
de aanpak van	the approach of	10757
de aanslagen van	the (terrorist) attacks of	1152
de actie van	the action of	19982
de afgelopen jaren	the past years	41346
de afgelopen maanden	the past months	17046
de Amerikaanse president	the American president	2542
de Amerikaanse regering	the American government	875
de andere kant	the other side	426100
de bacterie is	the bacteria is	45
de besten van	the best of	989
de bouw van	the construction of	37885
de buurt van	the neighborhood of	217614
de dag dat	the day that	194752
de discussie over	the discussion on	13560
de dood van	the death of	42027
de economie van	the economy of	3566
de eerste plaats	the first place	34783
de finale van	the finals of	103195
de foto van	the picture of	68385

de gevolgen van	the consequences of	30608
de halve finale	the semi-finals of	81520
de handen vol	hands full	1988
de hele dag	the whole day	1946081
de hele wereld	the whole world	249719
de helft van	half of	417353
de inval in	the invasion of	666
de jaren negentig	the nineties	2893
de jaren twintig	the twenties	387
de kans dat	the chance that	41581
de keuze van	the choice of	12330
de komende jaren	the coming years	48187
de komende vier	the next four	6470
de komst van	the arrival of	48370
de kwaliteit van	the quality of	45240
de laatste twintig	the last twenty	507
de markt is	the market is	6906
de mensen hier	the people here	11779
de mogelijkheid om	the possibility to	31704
de moord op	the murder on	24169
de nabijheid van	the proximity of	2100
de nationale ploeg	the national team	3493
de ontvangst van	the reception of	2864
de oorlog in	into the war	5740
de oorlog tegen	the war against	1573
de organisatie van	the organization of	24050
de ploeg van	the team of	6725
de politie had	the police had	957
de positie van	the position of	11485
de presentatie van	the presentation of	37769
de prijs van	the price of	59208
de rand van	the edge of	66531
de rechtbank in	the court in	17346
de rest van	the rest of	782197
de resultaten van	the results of	16923
de rol van	the role of	55053
de Russische president	the Russian president	1116
de sociale zekerheid	the social security	2577
de strijd tegen	the battle against	34540
de trainer van	the coach of	10399
de tweede helft	the second half	59770
de tweede plaats	the second place	11487
de tweede ronde	the second round	22796
de universiteit van	the university of	5810
de vader van	the father of	68116
de vleugels van	the wings of	2756
de vraag is	the question is	62321
de website van	the website of	107326
de woorden van	the words of	24221
de zoon van	the son of	28840
door het ministerie	by the ministry	704
door onze correspondent	by our correspondent	105
drie jaar geleden	three years ago	9994
een aantal weken	a couple of weeks	16249
een actie van	an action by	11974
een belangrijke rol	an important role	10667

een bezoek aan	a visit to	47188
een brief aan	a letter to	9675
een deel van	a part of	130958
een film van	a movie of	20778
een gesprek met	a conversation with	97516
een groot aantal	a large number	16102
een groot deel	a big part	49737
een half jaar	half a year	220540
een half uur	half an hour	1031387
een hoger niveau	a higher level	11297
een idee van	an idea of	9561
een jaar eerder	a year earlier	4725
een jaar geleden	a year ago	123936
een kans om	a chance to	17502
een kwart van	a quarter of	17557
een kwestie van	a question of	80116
een moment dat	a moment that	17900
een onderzoek naar	an investigation into	16947
een opkomst van	a rise of	1276
een paar dagen	a couple of days	399097
een paar jaar	a couple of years	155619
een paar weken	a couple of weeks	177710
een tentoonstelling van	an exhibition of	1062
een vorm van	a form of	43031
een winst van	a profit of	3386
een woordvoerder van	a spokesman for	2266
eerder deze week	previously this week	13746
einde van het	end of the	77611
elke keer weer	every time again	92297
en de manier	and the way	6222
en te weinig	and too little	18505
euro per maand	euro per month	47736
genieten van een	enjoying a	132200
ging het mis	it went wrong	64252
het afgelopen jaar	the past year	35986
het begin van	the beginning of	212462
het belang van	the importance of	51523
het bestuur van	the board of	17895
het centrum van	the center of	128503
het centrum voor	the center for	2134
het eerst sinds	for the first time since	91715
het eerste kwartaal	the first quarter	19414
het functioneren van	the functioning of	2732
het gebied van	the area of	128659
het gebruik van	the use of	63198
het gevoel van	the feeling of	29980
het ging om	it was about	15042
het herstel van	the recovery of	5251
het is geen	it is no	155183
het is niet	it is not	512730
het kader van	the framework of	128254
het laatste kwartier	the last quarter	4456
het ministerie van	the ministry of	19317
het moment dat	the moment that	158686
het najaar van	the fall of	2818
het plan van	the plan of	10331

het spel van	the game of	13382
het tweede kwartaal	the second quarter	14558
het vertrek van	the departure of	19881
het werk van	the work of	45904
het zoeken naar	the search for	27018
ik denk dat	I think that	1770156
in de aanloop	in the run-up	9802
in de auto	in the car	1520967
in de buurt	in the neighborhood/close	944910
in de eredivisie	in the premier division	39637
in de hoek	at the corner	53419
in de krant	in the newspaper	224065
in de lucht	in the air	320247
in de ogen	in the eyes	62735
in de omgeving	in the neighborhood	73136
in de partij	in the party	3097
in de politiek	in politics	47360
in de praktijk	in practice	75131
in de regio	in the region	175792
in de rij	in line	202447
in de stad	in the city	1106466
in de strijd	in the battle	52453
in de wereld	in the world	207594
in de zomer	during summer	335094
in de zorg	in health care	125713
in dit gebouw	in this building	2176
in een open	in an open	5010
in eigen land	in your own country	32385
in Europa zijn	being in Europe	2226
in handen van	in the hands of	22157
in het begin	at the beginning	133282
in het boek	in the book	51266
in het centrum	in the center	143835
in het kader	in the framework	78640
in het land	in the country	93987
in het noorden	in the north	58157
in het openbaar	in public	67556
in het verhaal	in the story	12894
in het ziekenhuis	in the hospital	350649
in ons land	in our country	41727
is een beetje	is a little	337220
is er ook	is there too	347370
is niet alleen	is not alone	75895
is niet nodig	is not necessary	68369
is nog steeds	still is	448018
is zo goed	is so good	77105
kans op een	chance of a	358386
kijken naar de	to look at the	105645
maakt niet uit	does not matter	1306258
maar het was	but it was	260391
met de mededeling	with the announcement	9055
met de speler	with the player	466
met een optie	with an option	2562
met een winst	with a profit	1453
moet ook nog	also has to	333650
na de oorlog	after the war	5851

na de pauze	after the break	28459
naar het buitenland	abroad	46869
net als hij	just like him	1766
niet al te	not too	195025
niet de enige	not the only one	432110
niet te veel	not too much	209056
niet te vergeten	not to forgot	116559
niets weten van	knowing nothing of	2986
nog een seizoen	yet another season	7423
nog niet bekend	not yet known	62111
nog ver weg	still far away	14766
nu is dat	now is that	19928
oktober vorig jaar	October last year	1739
om dit jaar	to ... this year	8176
om het leven	to the life	52336
om te buigen	in order to bend	2712
om te overleven	in order to survive	15291
ook wel eens	every now and then	332971
op dat moment	at that moment	103274
op de bank	on the couch	3286884
op de beurs	at the stock market	46130
op de dag	on the day	159050
op de eerste	at the first	126026
op de markt	on the market	176993
op de schouder	on the shoulder	5291
op de website	on the website	229875
op die manier	in this way	92344
op dit moment	at this moment	1037494
op een aanslag	at an attack	503
op het moment	at the moment	222221
op het nieuwe	at/on the new	34288
op het werk	at work	386510
op langere termijn	in the long run	3658
op te lossen	to solve	104502
op zijn bureau	at his desk	1171
op zijn minst	at least	27093
over de geschiedenis	about the history	8633
over de manier	about the way	4767
over de toekomst	about the future	43455
over het land	about the country	9189
paar jaar geleden	couple of years ago	57132
pas sinds kort	only recently	3790
raad van bestuur	board of directors	7008
sinds lange tijd	since a long time	26597
te beseffen dat	to realize that	16136
te staan in	to stand in	7740
te weten dat	to know that	106478
te zien zijn	to be visible	32458
terug te komen	to return	46567
tijdens de dictatuur	during the dictatorship	30
tot nu toe	until now	632637
twee weken op	two weeks on	6750
uit de grond	from the ground	31069
uit de ploeg	from the team	874
uit een boek	from a book	5952
uit eigen ervaring	from my own experience	8863

uit het Engels	from English	1175
uit te leggen	to explain	172789
uit te nodigen	to invite	16723
van de aandelen	of the shares	2190
van de aarde	from the earth	29344
van de beurs	from the stock market	8974
van de bevolking	from the population	18109
van de drie	out of three	33303
van de economie	of the economy	8738
van de families	of the families	263
van de gemeente	of the municipality	102650
van de jaren	of the years	15061
van de mogelijkheden	of the possibilities	4651
van de overheid	from the government	31026
van de tien	from the ten	11905
van de trainer	from the trainer	6949
van de twee	from the two	50064
van de volgende	from the next	17977
van de vorige	from the previous	43525
van de wereld	from/of the world	447158
van dit jaar	of this year	181143
van het aanbod	of the offer	2090
van het Britse	of the British	3003
van het kabinet	from the cabinet	18244
van het land	from/of the country	116573
van het museum	from/of the museum	4352
van het onderzoek	from/of the research	8406
van het publiek	from the public	10216
van het seizoen	from/of the season	202065
van onze redactie	from our editorial staff	848
van zijn proces	of his process	208
verdacht van fraude	suspected of fraud	1469
volgens het boekje	according to the rules	3106
voor de bescherming	for the protection	1022
voor eigen publiek	for your own audience	7444
voor het eerst	for the first time	1025405
voor het leven	for life	106138
voor iemand die	for someone who	89362
wel of niet	yes or no	286718
wordt beschuldigd van	is being accused of	1593
zo lang geleden	so long ago	40677

Table 1: Stimuli used in the eye-tracking experiment of Chapter 2. Translations in English are provided in the second column. Phrasal frequencies (freqABC) are listed in the righter-most column.

Appendix B

trigram	translation	condition	frequency ABC	frequency C
aan de beurt	turn	MWU	1743	7458
aan de prins	to the prince	Control	80	12030
aan het eind	at the end	MWU	6395	33356
aan het bed	at the bed	Control	140	24021
aan te passen	to adapt	MWU	2307	10335
aan te steken	to strike	Control	103	7385
de andere kant	the other side	MWU	8113	30948
de andere groep	the other group	Control	108	38917
de eerste plaats	the first place	MWU	5853	80379
de eerste vraag	the first question	Control	237	63831
de hele dag	the whole day	MWU	5243	112638
de hele weg	the whole way	Control	142	101363
een belangrijke rol	an important role	MWU	2875	33764
een belangrijke vorm	an important shape	Control	21	33661
een groot deel	a big part	MWU	5427	69946
een groot kind	a big child	Control	26	46095
een paar dagen	a couple of days	MWU	4809	56414
een paar miljoen	a couple of million	Control	159	58205
er is geen	there is no	MWU	5264	384494
er is wat	there is something	Control	101	525923
in de praktijk	in practice	MWU	5764	11017
in de uitspraak	in the statement	Control	79	9997
in dit geval	in this case	MWU	4333	55463
in dit verhaal	in this story	Control	240	34648
in het centrum	in the center	MWU	3284	15338
in het vertrek	in the room	Control	73	8615
is de kans	is the chance	MWU	1224	30148
is de druk	is the pressure	Control	78	33051
mee te maken	to experience	MWU	2548	143877
mee te komen	to join	Control	90	130534
met de auto	by car	MWU	1302	28882
met de partij	with the party	Control	117	35702
na te denken	to think about	MWU	2349	42304
na te vragen	to inquire	Control	28	47636
nog een keer	again	MWU	5217	91196
nog een week	another week	Control	324	76650

om te kijken	to watch	MWU	1987	50552
om te nemen	to take	Control	18	58350
op dat moment	at that moment	MWU	6441	48502
op dat idee	on that idea	Control	58	37396
op de markt	on the market	MWU	5900	24936
op de brief	on the letter	Control	159	17057
op korte termijn	at short-notice	MWU	2060	14004
op korte afstand	at short distance	Control	104	13658
op te lossen	to solve	MWU	2974	4409
op te drukken	to push up	Control	23	4308
van de bevolking	from the population	MWU	6022	20512
van de discussie	from the discussion	Control	319	17251
van het jaar	of the year	MWU	6150	294718
van het nu	from nowadays	Control	75	358059
voor de toekomst	for the future	MWU	1926	23744
voor de liefde	for the love	Control	127	19180

Table 1: Stimuli used in the experiment. On the left all frequent multi-word units are listed, directly followed by their matched control. Translations in English are provided in the second column. Phrasal frequencies (freqABC) and frequencies of the third word (freqC) are listed in the two righter-most columns next to the stimuli.

Bibliography

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., and Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PloS one*, 12(4):e0174623.

Arnon, I. and Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, 56(3):349–371.

Arnon, I. and Priva, U. C. (2014). Time and again: The changing effect of word and multi-word frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon*, 9(3):377–400.

Arnon, I. and Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122(3):292–305.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Baayen, H., Vasishth, S., Kliegl, R., and Bates, D. (2017a). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94:206–234.

Baayen, R. and Ramscar, M. (2015). Abstraction, storage and naive discriminative learning. *Handbook of Cognitive Linguistics*, 39:100–120.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3):436–461.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Baayen, R. H., Hendrix, P., and Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, 56(3):329–347.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438.

Baayen, R. H., Sering, T., Shaoul, C., and Milin, P. (2017b). Language comprehension as a multiple label classification problem.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016a). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, 31(1):106–128.

Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. N. (2016b). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. *arXiv preprint arXiv:1601.02043*.

Baisa, V., Michelfeit, J., Medved, M., and Jakubícek, M. (2016). European union language resources in sketch engine. In *LREC*.

Bannard, C. and Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19(3):241–248.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, 59(s1):1–26.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Ltd.,-1999.–1204 p.

Boersma, P. and Weenink, D. (2016). Praat: Doing phonetics by computer.[computer program]. version 6.0. 19.

Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Brink, D. v. d. and Hagoort, P. (2004). The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by erps. *Journal of Cognitive Neuroscience*, 16(6):1068–1084.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Carrol, G. and Conklin, K. (2015). Eye-tracking multi-word units: Some methodological questions. *Journal of Eye Movement Research*, 7(5).

Conklin, K. and Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61.

Connolly, J. F. and Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, 6(3):256–266.

Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes*, 13(1):21–58.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

Dąbrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing.

De Cat, C., Klepousniotou, E., and Baayen, R. H. (2015). Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. *Frontiers in psychology*, 6.

Dehaene, S. (2009). *Reading in the brain: The new science of how we read*. Penguin.

Dikker, S. and Pylkkänen, L. (2013). Predicting language: Meg evidence for lexical preactivation. *Brain and language*, 127(1):55–64.

Durrant, P. and Doherty, A. (2010). Are high-frequency collocations psychologically real? investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2):125–155.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1):29–62.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4):491–505.

Fisher, K., Bassok, M., and Osterhout, L. (2010). When two plus two does not equal four: Event-related potential responses to semantically incongrous arithmetic word problems. In *Proceedings of the Cognitive Science Society*, volume 32.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2):78–84.

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268.

Geeraert, K., Newman, J., and Baayen, R. H. (2017). Idiom variation: Experimental data and a blueprint of a computational model. *Topics in Cognitive Science*.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Green, C. (2017). Usage-based linguistics and the magic number four. *Cognitive Linguistics*, 28(2):209–237.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage*, 20:S18–S29.

Hagoort, P. and Brown, C. M. (2000). ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia*, 38(11):1518–1530.

Han, S. (2015). *Processing formulaic sequences by native and nonnative speakers of English: Evidence from reading aloud*. PhD thesis, Northern Arizona University.

Harley, T. A. (2013). *The psychology of language: From data to theory*. Psychology press.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., and Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, 30(4):1383–1400.

Hauk, O. and Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5):1090–1103.

Hendrix, P., Bolger, P., and Baayen, H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(1):128.

Holcomb, P. J. and Neville, H. J. (1991). Natural speech processing: An analysis using event-related brain potentials. *Psychobiology*, 19(4):286–300.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., and Suchomel, V. (2013). The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

Janssen, N. and Barber, H. A. (2012). Phrase frequency effects in language production. *PloS one*, 7(3):e33202.

Jiang, N. A. and Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91(3):433–445.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329.

Kaan, E. and Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of cognitive neuroscience*, 15(1):98–110.

Keuleers, E., Stevens, M., Mandera, P., and Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, 68(8):1665–1692.

King, J. W. and Kutas, M. (1995). Who did what and when? using word-and clause-level erps to monitor working memory usage in reading. *Journal of cognitive neuroscience*, 7(3):376–395.

Kircher, T. T., Brammer, M., Andreu, N. T., Williams, S. C., and McGuire, P. K. (2001). Engagement of right temporal cortex during processing of linguistic context. *Neuropsychologia*, 39(8):798–809.

Kluender, R. and Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, 5(2):196–214.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Kryuchkova, T., Tucker, B. V., Wurm, L. H., and Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. *Brain and Language*, 122(2):81–91.

Kuiper, K. (1996). Smooth talkers. *The linguistic performance of auctioneers and sportscasters. Mawah, NJ: Lawrence Erlabum Associates.*

Kuiper, K., Van Egmond, M.-E., Kempen, G., and Sprenger, S. (2007). Slipping on superlemmas: Multi-word lexical items in speech production. *The Mental Lexicon*, 2(3):313–357.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading polymorphemic dutch compounds: Toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):876.

Kutas, M. and Van Petten, C. (1994). Psycholinguistics electrified. *Handbook of psycholinguistics*, pages 83–143.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics:(de) constructing the n400. *Nature Reviews Neuroscience*, 9(12):920.

Laubrock, J., Kliegl, R., and Engbert, R. (2006). SWIFT explorations of age differences in eye movements during reading. *Neuroscience & Biobehavioral Reviews*, 30(6):872–884.

Lazaridou, A., Marelli, M., Zamparelli, R., and Baroni, M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526.

Lensink, S. E., Schiller, N. O., and Verhagen, A. (submitted). Old and young: How language experiences (do not) shape the reading of lexical bundles.

Lensink, S. E., Verdonschot, R., Schiller, N. O., and Tamaoka, K. (in preparation). Reading japanese lexical bundles.

Lensink, S. E., Verhagen, A., Schiller, N. O., and Baayen, R. H. (submitted). Keeping it apart: On using a discriminative approach to study the nature and processing of multi-word units.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology)*, 61(2):381–400.

Linke, M., Bröker, F., Ramscar, M., and Baayen, H. (2017). Are baboons learning "orthographic" representations? Probably not. *PloS one*, 12(8):e0183876.

Lõo, K., Järvikivi, J., and Baayen, R. H. (2017). Whole-word frequency and inflectional paradigm size facilitate estonian case-inflected noun processing. *Cognition*.

Lõo, K., Järvikivi, J., Tomaschek, F., Tucker, B. V., and Baayen, R. H. (2018). Production of estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Morphology*, pages 1–27.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Martín-Loeches, M., Muñoz, F., Casado, P., Melcon, A., and Fernández-Frías, C. (2005). Are the anterior negativities to grammatical violations indexing working memory? *Psychophysiology*, 42(5):508–519.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, 18(1):1–86.

McGowan, V. A. and Reichle, E. D. (2018). The "risky" reading strategy revisited: New simulations using ez reader. *Quarterly Journal of Experimental Psychology*, 71(1):179–189.

McWhinney, S. R., Tremblay, A., Chevalier, T. M., Lim, V. K., and Newman, A. J. (2016). Using cforest to analyze diffusion tensor imaging data: A study of white matter integrity in healthy aging. *Brain connectivity*, 6(10):747–758.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017). Discrimination in lexical decision. *PLoS One*, 12(2):e0171935.

Milin, P., Kuperman, V., Kostic, A., and Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in grammar: Form and acquisition*, pages 214–252.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *ACL*, pages 236–244.

Miwa, K., Libben, G., and Ikemoto, Y. (2017). Visual trimorphemic compound recognition in a morphographic script. *Language, Cognition and Neuroscience*, 32(1):1–20.

Mueller, J. L., Hahne, A., Fujii, Y., and Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, 17(8):1229–1244.

Müller, H. M., King, J. W., and Kutas, M. (1997). Event-related potentials elicited by spoken relative clauses. *Cognitive Brain Research*, 5(3):193–203.

Norris, D. and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

Oostdijk, N., Reynaert, M., Hoste, V., and Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential speech and language technology for Dutch*, pages 219–247. Springer.

Pawley, A. and Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, 191:225.

Penolazzi, B., Hauk, O., and Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, 74(3):374–388.

Pinker, S. and Ullman, M. T. (2002). The past and future of the past tense. *Trends in cognitive sciences*, 6(11):456–463.

Pollatsek, A., Hyönä, J., and Bertram, R. (2000). The role of morphological constituents in reading finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):820.

Pylkkänen, L., Feintuch, S., Hopkins, E., and Marantz, A. (2004). Neural correlates of the effects of morphological family frequency and family size: an meg study. *Cognition*, 91(3):B35–B45.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, 46(4):377–396.

Ramscar, M., Dye, M., and McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4):760–793.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, 6(1):5–42.

Ramscar, M. and Port, R. (2015). Categorization (without categories). In Dąbrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 75–99. De Gruyter, Berlin.

Ramscar, M. and 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., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Rayner, K., Castelhano, M. S., and Yang, J. (2009). Eye movements and the perceptual span in older and younger readers. *Psychology and aging*, 24(3):755.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., and Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and aging*, 21(3):448.

Rayner, K., Yang, J., Schuett, S., and Slattery, T. J. (2014). The effect of foveal and parafoveal masks on the eye movements of older and younger readers. *Psychology and aging*, 29(2):205.

Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

Reichle, E. D., Rayner, K., and Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the ez reader model. *Vision research*, 39(26):4403–4411.

Reichle, E. D., Rayner, K., and Pollatsek, A. (2012). Eye movements in reading versus nonreading tasks: Using ez reader to understand the role of word/stimulus familiarity. *Visual cognition*, 20(4-5):360–390.

Reifegerste, J., Meyer, A. S., and Zwitserlood, P. (2017). Inflectional complexity and experience affect plural processing in younger and older readers of dutch and german. *Language, Cognition and Neuroscience*, 32(4):471–487.

Rescorla, R. A., Wagner, A. R., et al. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99.

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., and Schlesewsky, M. (2007). To predict or not to predict: Influences of task and strategy on the processing of semantic relations. *Journal of Cognitive Neuroscience*, 19(8):1259–1274.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sereno, S. C., Brewer, C. C., and O'Donnell, P. J. (2003). Context effects in word recognition: Evidence for early interactive processing. *Psychological Science*, 14(4):328–333.

Sereno, S. C., Rayner, K., and Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport*, 9(10):2195–2200.

Shafto, M. A. and Tyler, L. K. (2014). Language in the aging brain: the network dynamics of cognitive decline and preservation. *Science*, 346(6209):583–587.

Shaoul, C., Baayen, R. H., and Westbury, C. F. (2014a). N-gram probability effects in a cloze task. *The Mental Lexicon*, 9(3):437–472.

Shaoul, C., Schilling, N., Bitschnau, S., Arppe, A., Hendrix, P., and Baayen, R. (2014b). NDL2: Naive discriminative learning. R package version 1.901, development version available upon request.

Shaoul, C. and Westbury, C. (2010). Exploring lexical co-occurrence space using hidex. *Behavior Research Methods*, 42(2):393–413.

Shaoul, C. and Westbury, C. (2011). Formulaic sequences: Do they exist and do they matter? *The Mental Lexicon*, 6(1):171–196.

Shaoul, C., Westbury, C. F., and Baayen, H. R. (2013). The subjective frequency of word n-grams. *Psihologija*, 46(4):497–537.

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.

Siyanova-Chanturia, A. (2015). On the "holistic" nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, 11(2):285–301.

Siyanova-Chanturia, A., Conklin, K., and Schmitt, N. (2011a). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

Siyanova-Chanturia, A., Conklin, K., and Van Heuven, W. J. (2011b). 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.

Snider, N. and Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. *Frequency Effects in Language Representation*, pages 127–163.

Sosa, A. V. and MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language*, 83(2):227–236.

Sprenger, S. A., Levelt, W. J., and Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, 54(2):161–184.

Steinhauer, K., Connolly, J. F., Stemmer, B., and Whitaker, H. (2008). Event-related potentials in the study of language. *Concise Encyclopedia of Brain and Language*, pages 91–104.

Steinhauer, K. and Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, 120(2):135–162.

Steinhauer, K., Drury, J. E., Portner, P., Walenski, M., and Ullman, M. T. (2010). Syntax, concepts, and logic in the temporal dynamics of language comprehension: Evidence from event-related potentials. *Neuropsychologia*, 48(6):1525–1542.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323.

Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests, and trees of york english: Was/were variation as a case study for statistical practice. *Language variation and change*, 24(2):135–178.

Tomasello, M. (2009). *Constructing a language*. Harvard university press.

Tremblay, A., Asp, E., Johnson, A., Migdal, M. Z., Bardouille, T., and Newman, A. J. (2016). What the networks tell us about serial and parallel processing. *The Mental Lexicon*, 11(1):115–160.

Tremblay, A., Baayen, H., Derwing, B., Libben, G., Tucker, B. V., and Westbury, C. (2012). Empirical evidence for an inflationist lexicon. *Yearbook Phraseoly*, 3:109–126.

Tremblay, A. and 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. *Perspectives on formulaic language: Acquisition and communication*, pages 151–173.

Tremblay, A., Derwing, B., and Libben, G. (2009). Are lexical bundles stored and processed as single units? *Working Papers of the Linguistics Circle*, 19(1):258.

Tremblay, A., Derwing, B., Libben, G., and Westbury, C. (2011). Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61(2):569–613.

Tremblay, A. and Tucker, B. V. (2011). The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, 6(2):302–324.

Underwood, G., Schmitt, N., and Galpin, A. (2004). The eyes have it. *Formulaic sequences: Acquisition, processing, and use*, 9:153.

Van Den Brink, D., Brown, C. M., and Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, 13(7):967–985.

Van Den Brink, D. and Hagoort, P. (2004). The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. *Journal of Cognitive Neuroscience*, 16(6):1068–1084.

Van Rij, J., Hollebrandse, B., and Hendriks, P. (2016). Children's eye gaze reveals their use of discourse context in object pronoun resolution. *Experimental Perspectives on Anaphora Resolution. Information Structural Evidence in the Race for Salience. Boston: De Gruyter*, pages 267–293.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., and Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, 22(8):1682–1700.

Vigneau, M., Beaucousin, V., Herve, P.-Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., and Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, 30(4):1414–1432.

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., and Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing?: Insights from a meta-analysis. *Neuroimage*, 54(1):577–593.

Vitu, F., McConkie, G. W., Kerr, P., and O'Regan, J. K. (2001). Fixation location effects on fixation durations during reading: An inverted optimal viewing position effect. *Vision Research*, 41(25-26):3513–3533.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. Technical report, STANFORD UNIV CA STANFORD ELECTRONICS LABS.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 90(3):669–692.

Winter, B. and Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, 1(1):7–18.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32:231–254.

Wurm, L. H. and Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72:37–48.

Yates, M., Friend, J., and Ploetz, D. M. (2008). The effect of phonological neighborhood density on eye movements during reading. *Cognition*, 107(2):685–692.

Nederlandse samenvatting

Veilingmeesters en sportcommentatoren staan bekend om de enorme snelheid waarmee ze praten over biedingen en ballen. Om dit te kunnen, maken ze gebruik van een beperkte set van zinnen en zegswijzen die ze als kant-en-klare brokstukken aan elkaar kunnen knopen. Toch zijn het niet alleen de veilingmeesters en de sportcommentatoren die vaak in vaste formules praten — we maken er allemaal gebruik van in ons dagelijks leven.

De schattingen lopen uiteen, maar over het algemeen wordt aangenomen dat zeker de helft van onze gesproken en geschreven taal bestaat uit formules, standaardzinnen, en vaak voorkomende combinaties van woorden. Sommige van deze combinaties zijn ondoorzichtig: De betekenis van het geheel is niet af te leiden uit de som van de betekenissen van de losse woorden. *Daar komt de aap uit de mouw* gaat niet letterlijk over apen die uit kledingstukken klimmen. Er zijn echter ook veelvoorkomende combinaties waarvan de betekenis transparant is: Als je weet wat *in*, *de*, en *auto*, betekenen, dan weet je ook wat *in de auto* betekent. Deze transparante combinaties worden 'lexicale bundels' genoemd in de literatuur, en vormen het onderwerp van onderzoek van deze dissertatie.

Omdat lexicale bundels veel vaker gebruikt worden dan op basis van kans verwacht kan worden, rijst de vraag waarom juist die combinaties door taalgebruikers gebruikt worden. Binnen de taalwetenschap bestaat de gebruiksgebaseerde (*usage-based*) taalbenadering, die stelt dat de cognitieve representatie van taal voortkomt uit de manier waarop taal gebruikt wordt. Vanuit deze benadering is het aannemelijk dat vaakvoorkomende combinaties van woorden door gebruik samensmelten en ingesleten raken als brokstukken van taal die als eenheden verwerkt worden, zonder dat de gebruiker keer op keer de losse woorden moet samenvoegen op basis van grammaticale regels. Dit proces van samensmelten en inslijten staat bekend als *chunking*, en is een welbekend proces in andere cognitieve taken. Het zorgt ervoor dat deze taken snel, soepel en foutloos uitgevoerd kunnen worden. Het onthouden van een telefoonnummer, bijvoorbeeld, gaat makkelijker als de cijfers in brokstukken van meerdere losse getallen worden geleerd.

Deze gebruiksgebaseerde benadering voorspelt dus dat hoogfrequente combinaties van woorden als eenheden worden verwerkt. De afgelopen jaren zijn er steeds meer experimentele studies uitgevoerd waarvan de resultaten lijken te bevestigen dat vaakvoorkomende combinaties van woorden als eenheden in de verwerking gebruikt worden. Zo is gebleken dat de frequenties van gehele combinaties een grote rol spelen in het voorspellen van de snelheid waarmee mensen lexicale bundels lezen en uitspreken, los van de frequenties van de losse woorden waaruit die combinaties bestaan. Er waren echter nog erg weinig studies gedaan naar andere talen dan het Engels; er was nog nauwelijks onderzoek gedaan naar de verwerking van gesproken lexicale bundels; geavanceerde statistische modellen om goed te begrijpen welke factoren een rol spelen tijdens de verwerking van lexicale bundels werden nog weinig toegepast; en er is ook nog maar weinig gebruik gemaakt van computationale modellen om betere inzichten te krijgen in het verwerkingsproces van lexicale bundels. De onderzoeken in deze dissertatie pogen deze hiaten op te vullen.

De belangrijkste vraag die deze dissertatie tracht te beantwoorden is hoe lexicale bundels verwerkt worden door lezers, luisteraars, en sprekers van het Nederlands. Daarbij wordt gebruik gemaakt van geavanceerde statistische technieken en een computationeel model dat een cognitief plausibel model van leren biedt. Door gebruik te maken van grote corpora van het Nederlands, was het mogelijk om vast te stellen welke woordcombinaties van drie woorden zeer vaak voorkomen. Na checks door twee onafhankelijke codeerders kon vervolgens bepaald worden welke hoogfrequente combinaties een transparante betekenis hebben, en dus als lexicale bundel van het Nederland beschouwd zouden kunnen worden. De dissertatie is verdeeld in drie delen, waarbij het eerste deel zich richt op het lezen van lexicale bundels door zowel jongere als oudere lezers, het tweede deel op het luisteren naar lexicale bundels, en het derde deel op het lezen en het uitspreken van lexicale bundels, en welke extra inzichten in deze processen een computationeel model kan toevoegen.

Hoofdstuk 2 richt zich op de vraag hoe jongere en oudere volwassenen lexicale bundels lezen, en of er verschillen zijn tussen de leeftijdsgroepen. De vraag of er verschillen bestaan tussen jongere en oudere lezers, komt voort uit een *usage-based* standpunt dat aanneemt dat de representatie van taal in het brein bij ieder individu het gevolg is van de individuele ervaring die deze persoon met taal heeft gehad. Daardoor heeft eenieder ook een unieke taalrepresentatie, omdat iedereen weer andere ervaring met taal opdoet. Aangenomen wordt dat een grotere blootstelling aan een bepaalde combinatie van woorden tot het versmelten van deze combinatie leidt, waarbij deze lexicale bundel steeds meer als eenheid in taalverwerking gebruikt wordt. Omdat ouderen een veel grotere ervaring met taal hebben dan jongeren, en dus ook veel vaker frequente lexicale bundels tegen zijn gekomen, volgt uit een *usage-based* benadering dat bij ouderen lexicale bundels anders gerepresenteerd zijn dan bij jongeren. Het kan zijn dat ouderen een sterkere en uitgebreidere representatie van lexicale bundels hebben door de grotere blootstelling, of juist minder gebruik maken van kant-en-klare brokstukken, omdat ze meer oefening hebben dan jongeren in het

ophalen van losse woorden uit het lexicon en het samenvoegen daarvan volgens de regels van de grammatica. Als er een verschil bestaat in de representatie van lexicale bundels, dan heeft dat zeer waarschijnlijk ook gevolgen voor de manier waarop deze verwerkt worden.

Door gebruik te maken van eye-tracking is in een experiment in kaart gebracht hoe zowel 60-plussers als twintigers hoogfrequente Nederlandse lexicale bundels lezen. We hebben gebruik gemaakt van statistisch modelleren om in staat te zijn de effecten van verschillende linguïstische eenheden van verschillende groottes in kaart te brengen — zijn het vooral de losse woorden die bijdragen aan hoe snel een combinatie van woorden wordt gelezen, of spelen ook de frequenties van combinaties van twee of zelfs drie woorden mee?

Het gekozen statistisch model, een zogenaamd *generalized additive mixed-effects model* of GAMM, is een regressiemodel dat in staat is om niet-lineaire relaties te modelleren, en daarbij bovendien ook rekening houdt met de individuele verschillen tussen proefpersonen die losstaan van de kenmerken van de lexicale bundels zelf, en het tijdsverloop door het experiment heen. Deze modelleertechniek maakt het mogelijk om te zien welke linguïstische factoren een rol spelen in de duur van verschillende onderdelen van het lezen, en wat voor vorm deze relatie heeft. Er zijn duidelijke aanwijzingen dat representaties van gehele lexicale bundels een rol spelen in lezen, en al in een vroeg stadium, aangezien er frequentie-effecten van trigrammen zijn gevonden in modellen van de duraties van al de eerste fixaties gemaakt op de bundels. Deze frequentie-effecten spelen samen met verschillende oogmotorische kenmerken, zoals de positie van een fixatie, een rol in de duur en het aantal fixaties.

Opvallend is dat deze frequentie-effecten een andere richting hebben dan verwacht: Hoe frequenter een lexicale bundel, hoe langer de eerste fixatie op deze bundel duurt. Dit is het *Inverted Frequency Effect* genoemd, en zou verklaard kunnen worden door ofwel 1) lexicale competitie die hoger is zodra er sprake is van hoogfrequente combinaties, waarbij een grotere competitie leidt tot een vertraagde en dus langere verwerking; 2) een leesstrategie die proefpersonen (on)bewust inzetten bij lexicale bundels versus 'gewone' combinaties van woorden of 3) een andere verwerking van lexicale bundels dan losse woorden, omdat de verwerking van lexicale bundels een ander en trager proces is dan het verwerken van losse woorden.

Er is geen enkel verschil gevonden in de manier waarop jongeren en ouderen lexicale bundels lezen. Dit is onverwacht vanuit een *usage-based* perspectief, waar de voorspelling zou zijn dat een verschil van dertig tot veertig jaar aan taalervaring een groot verschil in representaties in het lexicon tot gevolg zou moeten hebben, en dus ook op online taalverwerking. Het zou kunnen zijn dat taalrepresentaties bij jongvolwassenen al gestabiliseerd zijn en nog maar weinig veranderen in de jaren daarna. Het is ook mogelijk dat de stimuli gebruikt voor dit experiment niet optimaal waren, of dat er door toeval geen effect is gevonden — een 'false negative'. Hoe het ook zij, het zal interessant zijn om in toekomstige experimenten vast te stellen of een grotere taalervaring daadwerkelijk geen effect heeft op de verwerking van lexicale bundels, of dat

de *usage-based* benadering deels herzien zal moeten worden.

Hoe mensen gesproken lexicale bundels verwerken, is het thema van hoofdstuk 3. Er bestond nog vrijwel geen onderzoek naar de online verwerking van gesproken lexicale bundels, en dit hoofdstuk bespreekt een experiment waarin proefpersonen moesten luisteren naar allerlei frequente lexicale bundels zoals *aan de beurt* en een laagfrequente tegenhanger zoals *aan de prins*. Hoewel beide combinaties bestaan uit losse woorden die gematcht zijn op hun frequentie, en met dezelfde twee woorden beginnen, zijn er grote verschillen in de frasale frequenties: *Aan de beurt* komt veel vaker voor dan *aan de prins*. Wanneer een proefpersoon deze lexicale bundels hoort, zal zij pas aan het einde van het tweede woord doorhebben wat het laatste woord zou kunnen zijn — een verwachte continuatie die het einde van een hoogfrequente lexicale bundel vormt (*aan de beurt*) of juist een onverwacht, maar even frequent woord, *prins*, dat in combinatie met *aan de* weinig voorkomt. Deze twee condities zijn vervolgens met elkaar vergeleken door met machine learning de ERP-data te analyseren.

De analysetechniek die voor deze dataset is gebruikt, is een *conditional inference random forest* (CForest). CForests zijn een krachtig machine learning algoritme waarbij een grote groep van verschillende beslisbomen worden gegenereerd. Iedere afzonderlijke beslisboom is gebaseerd op een willekeurige subset van de data, en voor iedere splitsing in de beslisboom wordt steeds uit een willekeurige subset van predictoren bepaald welke predictor de beste tweedeling in de data maakt. Dit zorgt voor een grote variatie in de afzonderlijke beslisbomen, die samen een bos of 'forest' vormen. Random forests zijn in staat om niet-lineaire relaties in de data vast te leggen, en staan bekend om hun grote nauwkeurigheid en stabiele voorspellingen.

Naast deze voordelen van random forests, is een belangrijke reden om bij deze EEG-studie voor CForests te kiezen, dat CForests het mogelijk maken om de effecten te beoordelen van sterk aan elkaar gecorreleerde predictoren. De frequentie van de gehele lexicale bundel is vaak sterk gecorreleerd met de frequenties van de bigrammen en unigrammen waaruit deze is opgebouwd. In regressie-analyses is het daarom niet mogelijk om al deze predictoren tegelijkertijd in één model mee te nemen — terwijl het heel goed mogelijk is dat al deze eenheden in parallel een effect hebben op de verwerking van lexicale bundels. Bovendien is EEG-data afkomstig van verschillende electrodes niet onafhankelijk van elkaar — het signaal gemeten door een willekeurige electrode is sterk gecorreleerd met het signaal van aangrenzende electrodes.

In de EEG-data is een duidelijk verschil te zien tussen het signaal gegenereerd door frequente lexicale bundels, en het signaal gegenereerd door de controle-items. Er is een continu en vroegbeginnend negatief signaal dat bij de controle-items een nog negatievere voltage had. In het random forest model is de vorm van het verloop van de voltages gemodeleerd door het effect van de lengte van de stimuli, de kans dat een woord op de derde plek van een stimulus zou staan, de frequenties van de losse woorden, de bigrammen, en de trigrammen mee te nemen, rekening te houden met de status van een item (een lexicale bundel of controle-item), het tijdsverloop, en de electrode waar het sig-

naal gemeten is. Door naar een representatieve boom uit de random forest te kijken, is het mogelijk om een deel van het random forest model te doorgronden en hypotheses te vormen over hoe auditief gepresenteerde lexicale bundels verwerkt worden.

We stellen voor dat bij de verwerking van gesproken lexicale bundels drie stadia te onderscheiden zijn. Als eerste worden er voorspellingen gemaakt over wat er zou kunnen komen, terwijl er tegelijkertijd volop *bottom-up* verwerking is. Na deze eerste stappen worden mogelijke concurrende vormen actief onderdrukt, terwijl er een competitie ontstaat tussen andere mogelijke lexicale kandidaten. Deze competitie zorgt ervoor dat de verwerking van lexicale bundels met vele lexicale concurrenten moeilijker is voor het cognitieve systeem. In de derde en laatste fase vindt de lexicale integratie plaats van alle vrijgekomen informatie. In al deze drie fasen is duidelijk dat de frequenties van zowel enkele woorden, bigrammen, en de gehele trigram een rol spelen, vaak parallel aan elkaar.

In hoofdstuk 4, ten slotte, worden de eerste stappen gezet naar een beter begrip van lexicale toegang tot lexicale bundels. Hoewel er een groeiend aantal studies is dat frequentie-effecten voor vaakvoorkomende combinaties van woorden vindt, wat veel onderzoekers doet vermoeden dat deze bundels een cognitieve realiteit hebben, is het niet duidelijk hoe het brein toegang krijgt tot deze bundels. Om beter te kunnen begrijpen wat een lexicale bundel is, helpt het om expliciet in een computermodel vast te leggen hoe lexicale toegang zou kunnen verlopen, en dan te testen of predictoren uit een dergelijk model even goed of zelfs beter in staat zijn om experimentele data te beschrijven dan traditionele predictoren zoals frequenties.

We hebben twee experimenten uitgevoerd, een leesexperiment waarbij we data van oogbewegingen registreerden met behulp van eye-tracking, en een productie-experiment, waarbij we registreerden hoe snel proefpersonen begonnen met hardop voorlezen van lexicale bundels, en hoe lang ze erover deden om deze bundels helemaal uit te spreken. De data van beide experimenten zijn gemodelleerd met zowel traditionele predictoren zoals de frequentie van de lexicale bundels, als predictoren uit een computationeel model van lexicale toegang, de Naive Discriminative Learner (NDL), waarbij lexicale bundels expliciet in het model zijn opgenomen.

NDL is een eenvoudig neuraal netwerk dat uit slechts twee lagen bestaat, een inputlaag waar fonemen, letters, of losse woorden de *cues* vormen die verbonden zijn met de outputlaag, een set van *outcomes* of uitkomsten, in dit geval symbolische eenheden, *lexomes*. Deze lexomen wijzen naar de locatie van lexicale bundels in een semantische ruimte, en zijn stabiele eenheden die een connectie vormen tussen immer veranderende taalvormen en betekenissen. In een NDL netwerk zijn alle *cues* met alle uitkomsten verbonden, en worden connecties gevormd via de Rescorla-Wagner leerregels. Deze leerregels zijn erg succesvol gebleken in het modelleren van uiteenlopende gedragingen van dieren, en vormen daarmee een cognitief plausibel algoritme dat gebruikt kan worden om te modelleren hoe mensen hun linguïstische kennis in de loop der jaren

opbouwen.

Volgens het NDL-model vormt een taalgebruiker op basis van woord-*cues* verwachtingen over welke lexicale bundel hij kan verwachten. Behalve verwachtingen op basis van de input, heeft een taalgebruiker ook verwachtingen op basis van eerdere ervaringen, zodat een lexicale bundel die vaker is gebruikt, ook eerder verwacht wordt. Door verwachtingen op basis van zowel de input als eerdere ervaringen te combineren, en dat te vergelijken met de daadwerkelijke combinaties van woorden in het signaal, leert het systeem van zowel correcte als incorrecte voorspellingen. De kracht van NDL zit niet alleen in het feit dat het rekening houdt met woorden die vaak samen voorkomen, maar dat het ook inzichtelijk maakt hoe onderscheidend een *cue* is: het woordje *een* kan door vele andere woorden gevolgd worden, en is dus een slechte *cue*, terwijl *paarse* een sterke *cue* vormt voor *krokodil*.

Een getraind NDL-netwerk vormt een mathematische karakterisatie van de toestand van het lexicon. Uit een dergelijk netwerk kunnen allerlei predictoren worden gehaald, zoals hoe sterk de verwachting van een bepaalde lexicale bundel op voorhand al is (een predictor die sterk lijkt op een traditionele frequentiemaat), hoe sterk bepaalde losse woorden de verwachting opwekken van bepaalde lexicale bundels, en hoe makkelijk bepaalde lexicale bundels van elkaar te onderscheiden zijn. Uit hoofstuk 4 is gebleken dat deze NDL-predictoren beter in staat zijn om de experimentele data te beschrijven dan traditionele frequentiematen alleen, en bovendien meer inzichten verschaffen.

Lexicale toegang tot lexicale bundels vindt plaats vanuit zowel een *top-down* als een *bottom-up* proces, waarbij trigram-frequenties een grote rol spelen, en een grotere co-activatie van vergelijkbare items het uitspreken van lexicale bundels versnelt. Als alleen gebruik gemaakt wordt van frequentiematen, zouden *bottom-up* en *top-down* processen niet los van elkaar beschouwd kunnen worden. Daarnaast blijkt uit de eye-trackingdata dat lezers sneller lexicale bundels lezen als ze meer tijd besteden aan de *first pass*, de eerste keer dat ze een stuk tekst van links naar rechts lezen.

Door de hele dissertatie heen komt keer op keer naar voren dat eenheden groter dan het woord een rol spelen in lezen, luisteren en spreken in het Nederlands. Uit de data blijkt dat het tijdsverloop en de processen betrokken bij taalverwerking vrijwel hetzelfde zijn bij losse woorden en bij lexicale bundels, wat suggereert dat hoogfrequente lexicale bundels op een zelfde manier functioneren en gerepresenteerd zijn. Het lijkt er bovendien op dat vaakvoorkomende lexicale bundels niet alleen vanwege hun frequentie als eenheid in het lexicon functioneren — semantische eigenschappen van deze bundels spelen vermoedelijk ook een rol. Kijkende naar de items gebruikt in deze dissertatie, valt op dat dit voornamelijk discoursmarkeringen zijn zoals *ik denk dat*, *affordances* zoals *op de tafel* en complexe tijd- op ruimtemarkeringen zoals *op de dag* of *in het midden*. Hoewel deze items puur op frequentie van de trigrammen geselecteerd zijn, blijken ze in het algemeen een soort functionele eenheden te zijn. Deze items worden toevallig door meerdere woorden uitgedrukt in het Nederlands, maar worden in (sommige) morfologisch rijke talen uitgedrukt als

een enkel woord. Deze dissertatie laat hiermee zien dat eenheden van vorm en betekenis niet altijd overeen hoeven te komen wat wij wegens orthografische redenen als meerdere losse woorden beschouwen.

Toch betekent dit niet dat lexicale bundels ondoorzichtige brokstukken zijn: Ook de kleinere eenheden waaruit de lexicale bundels zijn opgebouwd, de bigrammen en unigrammen, spelen een rol in de verwerking, parallel aan de gehele lexicale bundels zelf. Dit laat zien dat, hoewel vaak voorkomende combinaties als brokstukken verwerkt worden, het taalsysteem deze brokstukken ook nog steeds opbreekt in kleinere delen, die ieder op zich ook van belang zijn in de verwerking. Dit is aanvullend bewijs voor een model van taalverwerking waarin meerdere eenheden van verschillende groottes parallel worden verwerkt.

De belangrijkste toevoegingen aan bestaand onderzoek zijn a) de focus op geavanceerde statistische modellen die subtiele patronen van verwerking aan het licht kunnen brengen; b) de eerste data die laten zien hoe oudere volwassenen lexicale bundels verwerken; c) een uitgebreidere analyse van de manier waarop gesproken lexicale bundels online verwerkt worden, en d) het inzetten van een computationeel model van lexicale toegang, waarin lexicale bundels als eenheden zijn opgenomen — dit maakt het mogelijk om lexicale toegang tot lexicale bundels in subprocessen op te delen en daardoor beter te begrijpen, en daarmee ook de status van lexicale bundels in het mentale lexicon te beschrijven.

About the author

Saskia Lensink was born on May 7th 1989 in Woudenberg, the Netherlands. She moved to Leiden to study linguistics, where she obtained her master's degree in 2014. She received a scholarship from the Landelijke Onderzoeksschool Taalwetenschap (LOT; the national research school in linguistics) to conduct her PhD on the on-line processing of lexical bundles at Leiden University. She furthermore received a scholarship from the Linguistic Society of America to attend the LSA Summer Institute at the University of Chicago in 2015. During her PhD, she was involved in several projects outside of her thesis topic, including studies on morphological priming, Spanish L2 learners, and computational stylometry. After finishing her PhD work, Saskia continued her career as a data scientist, where she continues working on statistical modeling, modeling language, and trying to make the world a better place. She is currently working at the Netherlands Organisation for Applied Scientific Research, TNO.