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-

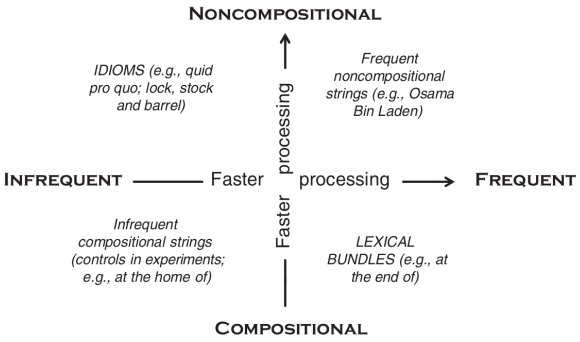

Figure 1.1: A typology of multi-word units. This thesis focuses on the units from the lower right corner, lexical bundles. Figure taken from Wray (2012, p. 241).

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.

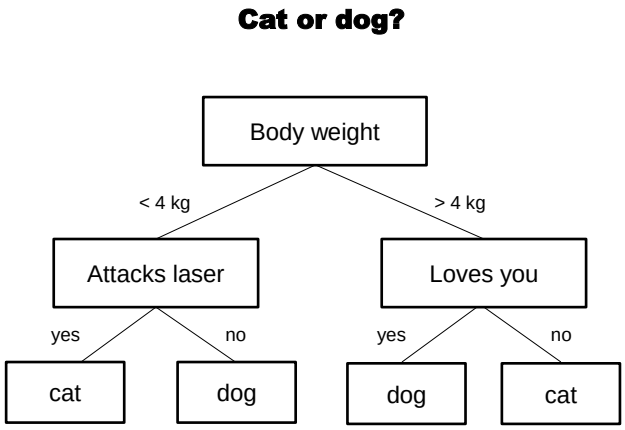

Figure 1.2: A decision tree that helps to distinguish between human's favorite pets, cats and dogs.

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.

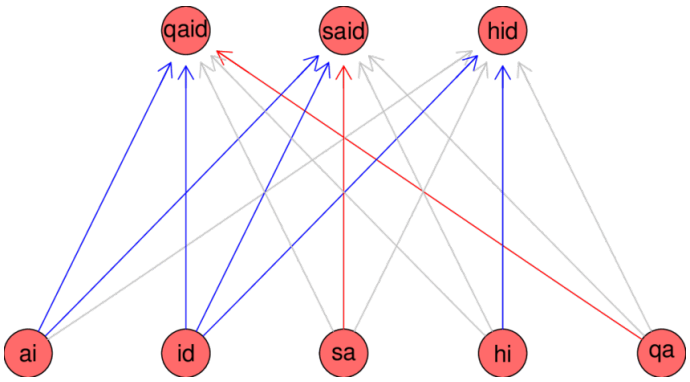

Figure 1.3: An NDL model with five digraphs as cues, and three lexemes as outcomes. Figure taken from Baayen and Ramscar (2015).

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.