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.