Processing Lexical Bundles

Lensink, S.E.

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.

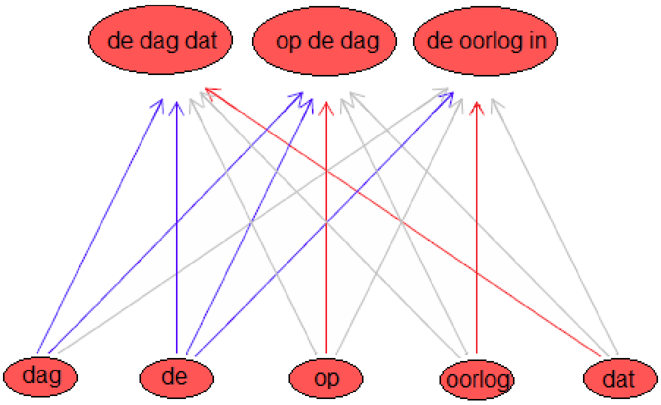

Figure 4.1: Part of the Rescorla-Wagner network used in this study, where the cues are formed by single words and the outcomes by the word trigrams used in the two experimental studies. Each cue is connected to all outcomes, and vice versa all outcomes are connected to each and every single cue. Red 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

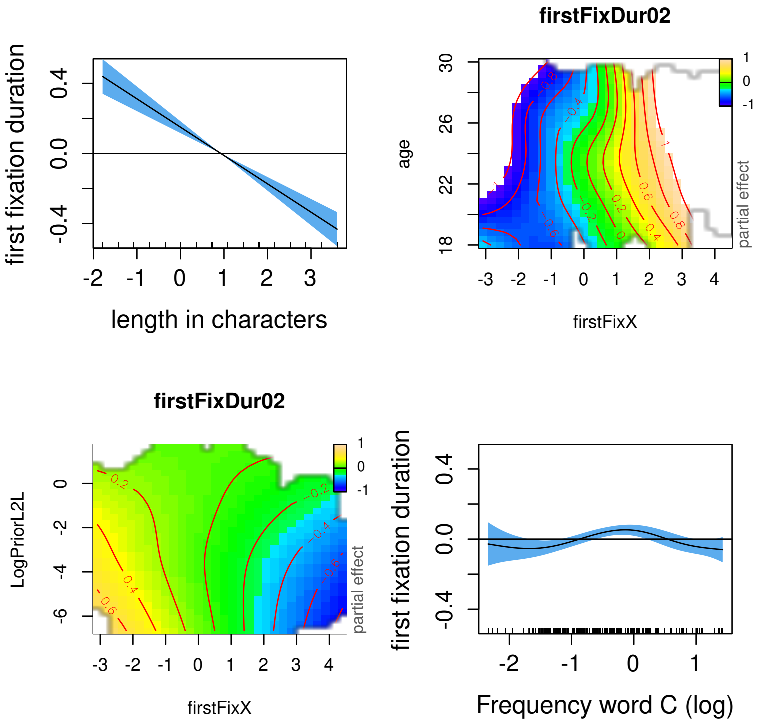

Figure 4.2: Partial effects of the model of the First Fixation Durations. The panel on the top left shows the effect of the lenght of the trigram, the panel on the top right shows the interaction of participant age and the horizontal location of the first fixation. The bottom left panel shows the interaction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

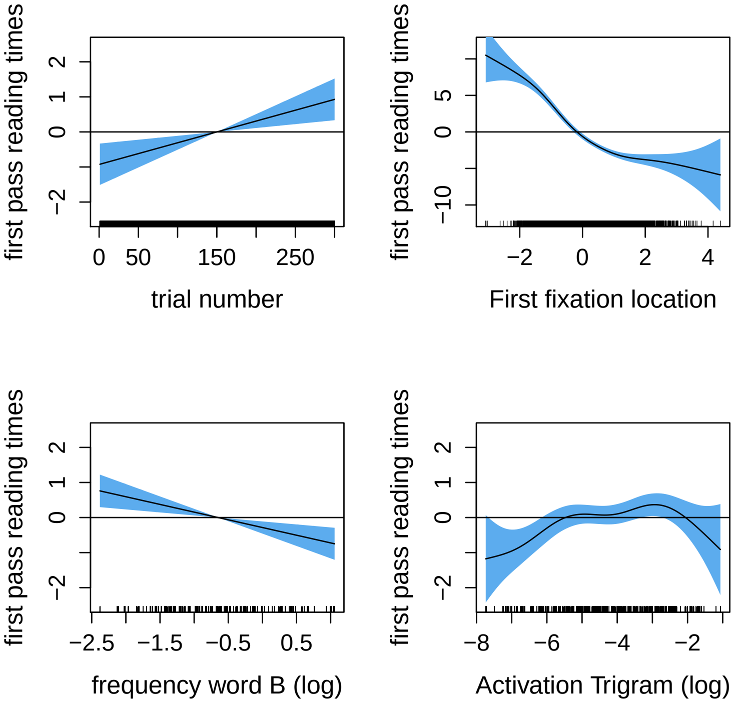

Figure 4.3: Partial effects of the model of the First Pass Reading Times. The top two panels show the effects of the trial number — reflecting the temporal position in the experiment — and the position of the first fixation. The bottom two panels show the effects of the second word frequencies and the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

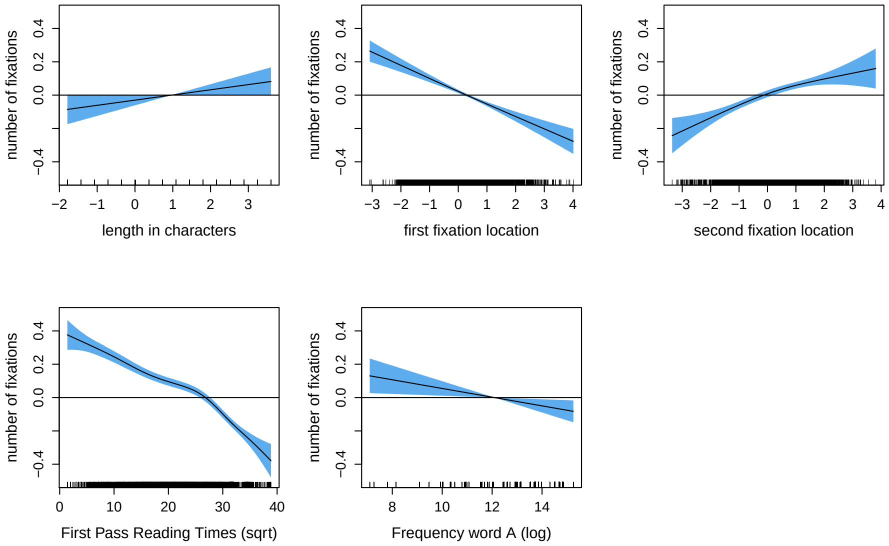

Figure 4.4: Partial effects of the model of the number of fixations. The top panels show the effects of the length of the trigram, and the horizontal locations of the first and second fixation. The bottom panels show the effects of the duration of the first pass reading times, and the frequency of the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

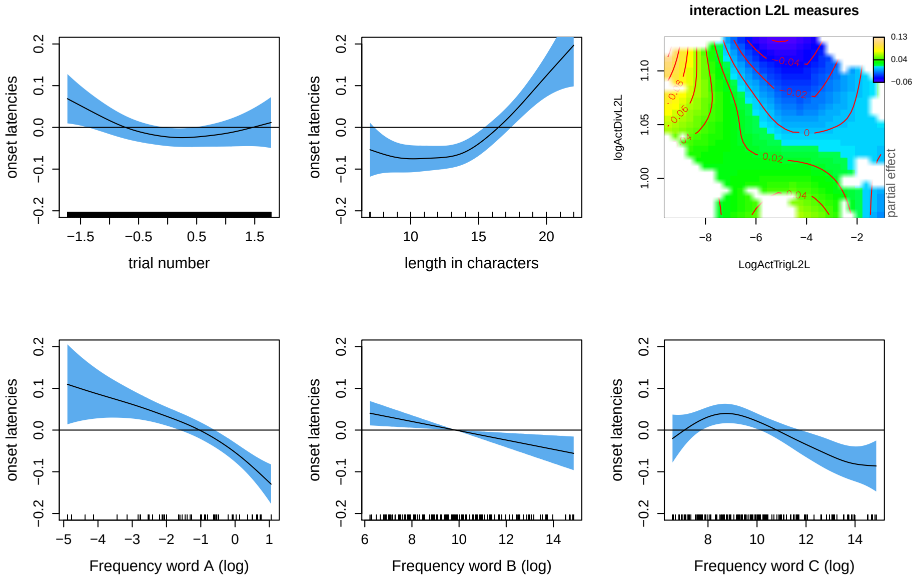

Figure 4.5: Partial effects of the model of the Onset Latencies of the production data. The first two top panels show the effects of trial number and length of the trigram. The panel at the top right shows the interaction of the NDL trigram activation (logActTrig) and the NDL trigram diversity (logActDiv).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

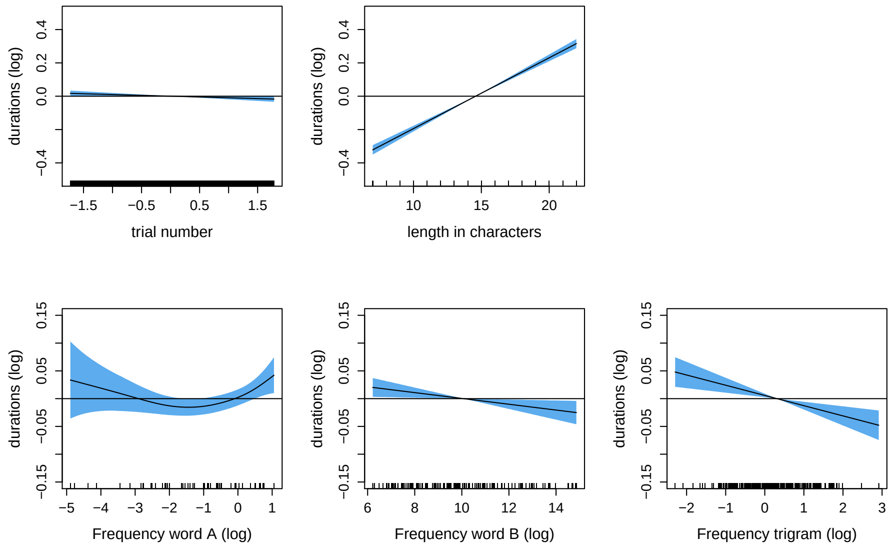

Figure 4.6: Partial effects of the model of the production durations. The top two panels show the effects of trial number and the length of the trigram. The bottom panels show the effects of the first word frequencies, the second word frequencies, and the trigram frequencies.

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.