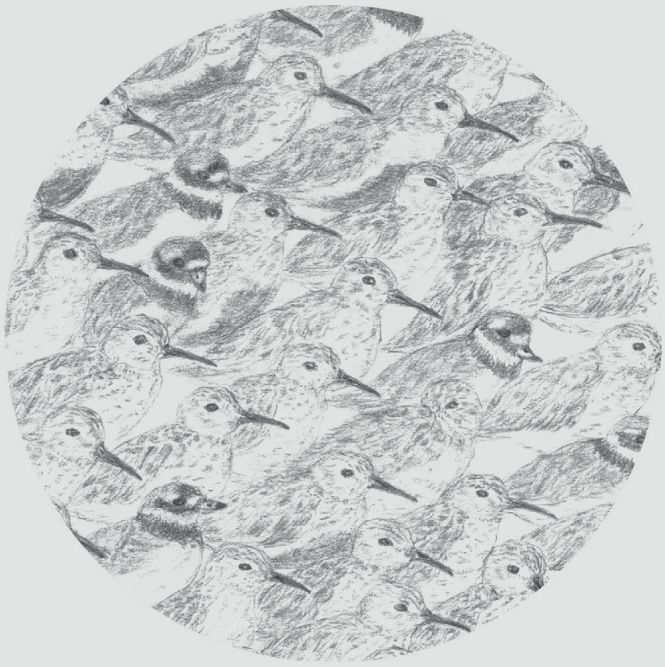


Words of a Feather Flock Together

The Role of Morphology in Human Auditory Word Recognition Models

HANNO MÜLLER

Centre for Language Studies

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Words of a Feather Flock Together

The Role of Morphology in Human Auditory Word Recognition Models

Hanno Maximilian Müller

The research presented in this thesis was funded by the Deutsche Forschungsgemeinschaft (Research Unit FOR2373 'Spoken Morphology', grant PL 151/7-2 'Central project' to Ingo Plag, and grant ER 574/1-1 'Dutch morphologically complex words: The role of morphology in speech production and comprehension' to Mirjam Ernestus, Louis ten Bosch, and Ingo Plag).

Words of a Feather Flock Together - The Role of Morphology in Human Auditory Word Recognition Models
Hanno Maximilian Müller

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Guus Gijben
Cover: Jenny Otto; drawing: Klaudia Müller
all drawing were made by Klaudia Müller and prepared for print by Jenny Otto
Printing: DPN Rikken/Pumbo

ISBN: 9789465152059
DOI: 10.54195/9789465152059
Free download at: https://doi.org/10.54195/9789465152059

RADBOUD
UNIVERSITY
PRESS

Words of a Feather Flock Together

The Role of Morphology in Human Auditory Word Recognition Models

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op
donderdag 26 maart 2026
om 10.30 uur precies

door

Hanno Maximilian Müller
geboren op 3 oktober 1991
te Keulen (Duitsland)

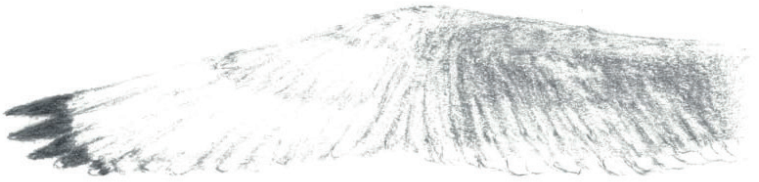
Promotor:
Prof. dr. M.T.C. Ernestus

Copromotor:
Dr. L.F.M. ten Bosch

Manuscriptcommissie:
Prof. dr. M.A. Larson
Prof. dr. I. Plag (Heinrich-Heine-Universität Düsseldorf, Duitsland)
Prof. dr. R.H. Baayen (Eberhard-Karls-Universität Tübingen, Duitsland)
Prof. dr. S. Arndt-Lappe (Universität Trier, Duitsland)
Dr. S.L. Frank

Es bleibt dabei:
die Zeitfolge ist das Gebiete des Dichters,
so wie der Raum das Gebiete des Malers.

Gotthold Ephraim Lessing, 1766



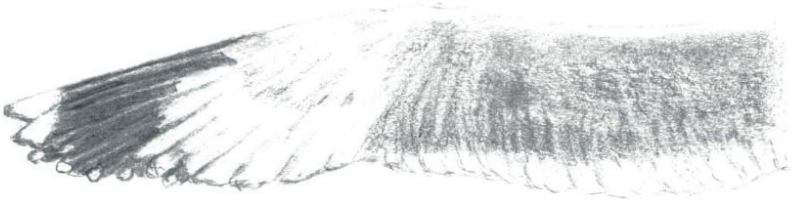


Table of contents

Chapter 1

General Introduction

Understanding spoken words is prerequisite for oral communication. We comprehend spoken words when chatting with our friends, maintaining social relationships. We process spoken words effortlessly when discussing ideas or concepts such as black holes. Listening to the news on radio or television, we comprehend spoken words, forming our perspective on what is going on in the world. The bottom line is that a tremendous number of information exchanges are based on the human ability to understand spoken words. This raises the question of how humans comprehend spoken words. This question is not only challenging because there is so much variation in how a word may sound (depending on the speaker, the speaking situation, the surrounding words, etc.), but also languages have thousands and thousands of words the listener has to choose from.

In many languages, collections of words share meaningful parts with each other. For instance, the words *view* and *viewer* share the part *view*, which in both words has a similar meaning. Such parts of words are called morphemes, and words consisting of more than one morpheme are called morphologically complex. Common morphologically complex words are prefixed and suffixed words. These are words that start and end, respectively, with a meaningful part that cannot form a word of its own. For instance, *-er* is a suffix in the word *viewer*. Examples of English words with prefixes are *interview*, *preview,* and *review*. The prefixes in these words are attached to the morpheme *view*, which is therefore called the words' base. Complex words can consist of a base and both a prefix and a suffix (e.g., *interviewer, previewing, reviews*). Finally, words can consist of several sequences of affixes (suffixes and prefixes) and bases, and can, consequently, be relatively long. An example is the Dutch word *aansprakelijkheidsverzekering* 'liability insurance', consisting of eight morphemes (i.e., *aan-, spraak, -elijk, -heid, -s, ver-, zeker, -ing*). Languages differ in how many affixes they have and in their constraints on how morphologically complex words can be formed. As a consequence, some languages, such as Estonian, have a rich (complex) morphological system, while others, for instance, German, Dutch, and English, have a less complex morphological system.

This thesis adds to the research on how listeners identify words, taking into account that a language contains both morphologically simple (consisting of only one morpheme) and morphologically complex words. On the one hand, it therefore extends the research on how listeners identify words (i.e., the field of spoken word recognition, also called auditory word recognition) by

taking into account morphologically complex words. On the other hand, it extends the research on the recognition of morphologically complex words by focussing on spoken word recognition (i.e., listening) instead of on visual (or written) word recognition (i.e., reading), which is the field in which the recognition of morphologically complex words is typically investigated (see Amenta & Crepaldi, 2012 for an overview). It is not self-evident that the language users recognise (i.e., process) morphologically complex words in the same way during reading and listening. Written words are presented at once and many words can be largely perceived in an instance due to parafoveal preview (Schotter, Angele, & Rayner, 2012). One gaze fixation spans multiple characters and often also multiple words. In contrast, spoken words unfold over time and therefore are perceived (if not processed) bit by bit.

More specifically, I will investigate a) whether the morphological structure of a spoken word influences how listeners recognise this word; b) which cognitive mechanisms underly effects of morphological structure in spoken word recognition; and c) which insights can be gained by investigating the recognition of spoken morphologically complex words with the research method of computational modelling. Computational modelling refers to the formal description of theories – in this thesis, theories of word recognition. Because this description is precise and complete, it allows the theory to make detailed predictions about how easily language users recognise a given word. These predictions can be compared with experimental data, for instance, with response times in listening experiments. This way, the theory, and therefore each of its assumptions, can be directly tested. The knowledge that is generated by this thesis' studies will help to build better theories and computational models of human spoken word recognition.

Spoken words are normally encountered in continuous speech. Nevertheless, most studies investigate how language users recognise words that were written or uttered without any context, that is, in isolation. This practice rests on the theoretical assumption that isolated words allow for the investigation of core recognition mechanisms without the additional variability introduced by sentence context, discourse structure, or pragmatic inference. Experimental paradigms presenting words in isolation are therefore used to capture fundamental properties of meaning access that also operate during continuous speech, even though additional processes come into play in speech recognition. Following this practice, this thesis will focus on how listeners recognise words that were uttered in isolation.

In what follows, in Chapter 1.1, I will first describe previous literature on the role of morphology in written word recognition – and if applicable in spoken word recognition –, focussing on word and morpheme frequency effects as well as the morphological family size effect. In Chapter 1.2, I will discuss theories that can account for effects of morphological structure in written word recognition, each adopting a systematically different approach. It is unclear whether these written word theories can account for effects of morphological structure in spoken word recognition and whether one theory is more plausible than another, a topic that I will investigate in this thesis. To this end, specific research methods are necessary, which I will discuss in Chapter 1.3. In Chapter 1.4, I will formulate the research questions of this thesis.

1.1 Experimental evidence

Over the last decades, a great body of studies investigated the recognition of morphologically complex words, particularly in the context of reading. Many of these studies focused on the question to what extent the frequency of occurrence of the words and of their morphemes determine how easily they are recognised by language users. Unless stated otherwise, in this thesis, the word *frequency* refers to token frequency, that is, the number of times a word form or morpheme occurs in a language (text) corpus. For instance, in the sentence 'The cat chased the cat', the words *the* and *cat* occur twice and thus each has a token frequency of two, whereas the word *chased* has a token frequency of one. Token frequency is taken to reflect the likelihood with which language users encounter a form. When frequency measures refer to other notions, such as type frequency (i.e., number of unique word types, in which a given morpheme occurs), this is explicitly indicated.

In what follows, I will focus on the often-replicated frequency effects that have shaped theories about the processing of morphologically complex words. That is, I will discuss morpheme frequency effects and morphological family size effects.

1.1.1 Morpheme frequency effects

Several studies have shown that the more frequent the morphemes in a word are, the faster this word is recognised. For instance, in a visual lexical decision experiment, Taft (1979) presented participants with inflected words that were matched on the frequency of their surface form (e.g., *likes*) but differed in their

stem frequency (i.e., the summed frequencies of other inflectional variants in the paradigm, e.g., *liked, liking*). He found that higher stem frequency words are responded to more quickly than lower stem frequency words. This phenomenon is referred to as stem frequency effect (e.g., Taft, 1979), root frequency effect (e.g., Meunier & Segui, 1999), or base frequency effect (e.g., Taft & Ardasinski, 2006). I will henceforth use the term *root frequency effect*. The root frequency effect was obtained by various researchers for various languages including Dutch (e.g., Baayen et al., 1997), French (e.g., Meunier & Segui, 1999), German (e.g., Kresse et al., 2012), Italian (e.g., Burani & Thornton, 2003), and Finnish (e.g., Lehtonen et al., 2007).

Although the root frequency effect appears to be quite robust, findings for the root frequency effect were also inconclusive. For instance, there are indications that root frequency only affects the recognition times of words with a low surface form frequency (e.g., Bradley, 1979; Burani & Caramazza, 1987) and may even inhibit the recognition of high surface form frequency words (Baayen, Wurm, & Aycock, 2007), although the facilitative root frequency effect has been reported for high surface form frequency words, too (Colé et al., 1989).

In spoken word recognition, the literature on the root frequency is less abundant. There are indications that root frequency exerts a facilitative effect on the recognition of Danish spoken words, although the root frequency effect only consistently emerges for low surface form frequency words (Winther Balling & Baayen, 2008, 2012). This finding contrasts with findings from French, where root frequency appears to facilitate the recognition of spoken high surface form frequency words (Meunier & Segui, 1999, Experiment 2). For French words with low surface form frequency, root frequency also has a facilitative effect, but only when controlling for the number of cohort members of the presented words (e.g., *viewable* and *viewability* are cohort members of *view*) and for the surface form frequencies of the presented words' morphological continuations (Meunier & Segui, 1999, Experiment 3).

1.1.2 Family size effects

Another finding that suggests that a word's morphological structure affects how this word is recognised is the so-called *morphological family size effect*. A word's morphological family size is the number of words (i.e., type count) that contain the word's root as a constituent. For instance, morphological family members of the English word *review* are, among others, *overview, preview, reviewed, reviews, view, viewed,* and *viewing*. In visual lexical decision experiments, words

are recognised more quickly, the larger their morphological families are (e.g., Schreuder & Baayen, 1997; de Jong, Schreuder, & Baayen, 2000; Juhasz & Berkowitz, 2011). This family size effect has been reported for different Western European languages such as Dutch (e.g., Moscoso del Prado Martín et al., 2004, 2005; Mulder, Dijkstra, & Baayen, 2015; Perdijk et al., 2012), English (Baayen et al., 2011; Dijkstra et al., 2005; Kuperman et al., 2009), and German (Lüdeling & de Jong, 2002), for Semitic languages such as Hebrew (Moscoso del Prado Martín et al., 2005) and Arabic (Boudelaa & Marslen-Wilson, 2011), for Mandarin (Feldman & Siok, 1997), and for Finnish (e.g., Moscoso del Prado Martín et al., 2004; Kuperman, Bertram, & Baayen, 2008). As the family size effect shows up in languages with different morphological systems, it can be assumed that this effect taps into a universal processing principle.

Previous literature suggests that the family size effect is especially driven by family members sharing a close *semantic* relationship with the presented word. The semantic nature of the family size effect was demonstrated, among others, in Finnish, where morphological family members can be divided into those that have a semantic relationship with the word presented and those that do not. For instance, the word *työ* 'work' has the semantically related family member *työehsopotimus* 'wage rate treaty' and the semantically unrelated family member *urotyö* 'heroic deed'. Moscoso del Prado Martín et al. (2004) conducted a Finnish visual lexical decision experiment (see Chapter 1.3 for more information about this type of experiment) and for predicting the response times (RTs), they tested a family size measure that is solely based on semantically related family members and a measure that is based on all family members. While the former measure significantly predicts the RTs, the latter measure does not. In Dutch, Bertram et al. (2000, Experiment 6) predicted the RTs of a visual lexical decision experiment with two different family size measures too. While the first measure is based on words that are semantically transparent, the second measure was based on all family members. For instance, given the presented word *gemeenheid* 'meanness', *gemeente* 'municipality' is considered to have a semantically opaque relationship with this word. The results suggest that only the former family size measure significantly predicts the RTs, whereas the latter measure does not. Similar findings were obtained by Schreuder and Baayen (1997).

Regarding orthographic relationships between the presented word and its family members, previous literature suggests that this relationship does not affect the family size effect. For instance, de Jong (2002) found that Dutch

irregular past participles elicit the facilitative family size effect in visual lexical decision, although in these past participles the root is differently spelled than in the citation form, which is the infinitive in Dutch. For instance, in *gedacht* 'thought', the root is spelled as *dacht* whereas the root is spelled *denk* in the citation form *denken* 'to think', in this verb's present tense forms, and in many other family members such as *denker* 'thinker' or *bedenkelijk* 'worrying'. These results were replicated by Bertram et al. (2000). Moscoso del Prado Martín et al. (2005) obtained similar results for Hebrew.

Whether the family size effect plays a role both in written and in spoken word recognition, is an unanswered question. The literature provides inconsistent findings with respect to the status of the family size effect in spoken word recognition. One study obtained the effect (Winther Balling & Baayen, 2012). Another study found that larger family sizes elicit slower responses (Winther Balling & Baayen, 2008). A third study failed to find an effect of family size at all (Baayen, Wurm, & Aycock, 2007).

1.2 Theories of morphological processing

Theories about morphological processing in word recognition can roughly be categorised according to how they represent linguistic units such as morphemes or words. Localist theories (e.g., Morton, 1969; Luce et al., 2000; Jackendoff & Audring, 2020; ten Bosch, Boves, & Ernestus, 2022; Taft, 2023) represent linguistic units with symbolic representations. For illustration, a localist model for Dutch could be based on a unique representation for each Dutch word. As soon as a representation is determined to match the input signal (i.e., an orthographic sequence or an audio signal) better than competing representations, the meaning associated with the input matching representation becomes available. Within the localist paradigm, theories can further be categorised according to whether they assume morpheme representations in the lexicon (e.g., Cutler, Hawkins, & Gilligan, 1985; Cutler & Norris, 1988; Diependaele, Sandra, & Grainger, 2009; Taft, 1979, 2003) or not (e.g., Adelman, 2011; Butterworth, 1983; Davis, 2010; ten Bosch, Boves, and Ernestus, 2022).

According to distributional–connectionist theories (Chuang & Baayen, 2021; Magnuson et al., 2018), neither morpheme nor word representations (or other intermediate linguistic units) have mental representations. Instead, the word

recognition process is conceptualised as the mapping of an incoming signal (i.e., an orthographic sequence or an audio signal) directly onto a corresponding meaning. This mapping is achieved by passing activation through a network of connections with as input form cues (e.g., letter n-grams or acoustic features) and as output abstract meaning representations. For establishing the association weights, the model first needs to be trained with supervised learning, that is, the model is presented with form cues and corresponding meanings. Figure 1. 1 schematically illustrates such a model training.

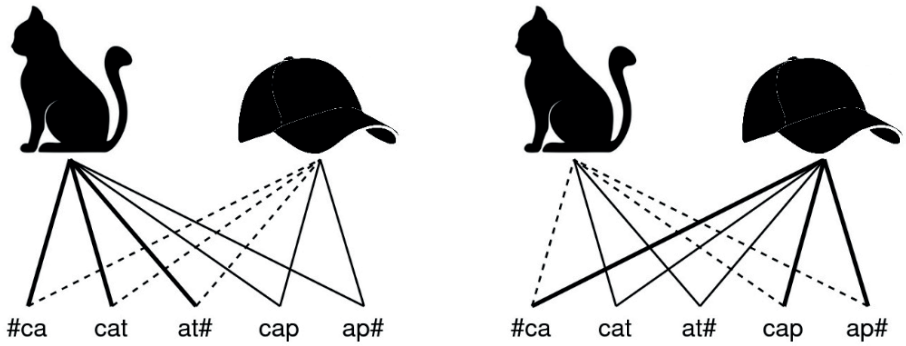


Figure 1.1 Schematic illustration of a Discriminative Learning Model (e.g., Chuang & Baayen, 2021) with letter trigrams as inputs and abstract semantic representations as outputs. Here, the # denotes a start-of-word or end-of-word symbol. During training, when presented with the word *cat* (left panel), associations between the trigrams of *cat* and the corresponding meaning of *cat* become stronger (thick solid lines), whereas associations between the trigrams of *cat* and other meanings representations become weaker (dashed lines). And vice versa, when presented with the word *cap* (right panel), this word's trigrams and its meaning become more strongly associated, whereas associations between these trigrams and other meanings are weakened.

In what follows, I will discuss how localist and distributional–connectionist theories account for the root frequency effect and the family size effect. To the best of my knowledge, neither localist nor distributional–connectionist theories have been implemented to directly simulate the root frequency effect. However, there are computational implementations of localist models that are grounded in the assumption that the root frequency effect indicates that word recognition involves access to morpheme representations. Regarding the family size effect, both localist and distributional–connectionist models have been implemented to successfully simulate the family size effect in reading.

1.2.1 Localist theories in light of frequency and family size effects

Localist explanations of the root frequency effect assume that morpheme representations in the mental lexicon are connected to the representations of the words that contain them. A higher-frequency root becomes activated more quickly, which in turn activates the complex words that include it, thereby facilitating their recognition (e.g., Meunier & Segui, 1999; Taft, 1979, 1994, 2023). Within this framework, the root frequency effect is taken as evidence that access to morpheme representations drives the recognition of morphologically complex words.

However, this view immediately raises tension with the surface form frequency effect: while root frequency effects imply that recognition is governed by root representations (e.g., Forster, 1976; Morton, 1979; Norris, 2006), surface form effects imply that recognition is governed by word-level representations (e.g., Brysbaert, Mandera, & Keuleers, 2018). To reconcile these apparently competing findings, hybrid localist models (e.g., Baayen, Dijkstra, & Schreuder, 1997; Diependaele, Sandra, & Grainger, 2009; Schreuder & Baayen, 1995) posit that roots and surface forms can both govern recognition, and their relative contributions may shift depending on factors such as suffix type (e.g., Baayen, Dijkstra, & Schreuder, 1997), the number of the presented word's cohort members and these members' form frequencies (Meunier & Segui, 1999), or the ratio between root and surface form frequencies (e.g., Hay, 2001).

Regarding the morphological family size effect, a localist explanatory account has been put forward in form of the *Morphological Family Resonance Model* (MFRM; de Jong, Schreuder, & Baayen, 2003). According to the MFRM, the word recognition process involves semantic representations (Schreuder & Baayen, 1995) and lemma representations (Levelt, 1989). In this model, semantic representations coincide with morphemes. That is, there are semantic representations for, among others, words (e.g., *book*), numerus (e.g., *PLURAL*), tense (e.g., *PAST TENSE*), or whether a word is a diminutive form (e.g., *DIMINUTIVE*). Because semantic representations represent morphemes, I will henceforth refer to these representations with morpheme representations. In the MFRM, the presented word's lemma representation spreads activation to its connected morpheme representations (see Figure 1.2, 1st cycle). Next, morpheme representations spread activation to their connected lemma representations, including the presented word's lemma representation (see Figure 1.2, 2nd cycle). The more family members a lemma has, the more family members' representations become activated via (the) shared root morpheme

representation(s). Then, the family members' representations activate the presented word's root morpheme representation (Figure 1.2, 3rd cycle), which will then activate the presented word's lemma representation. As a result, words with more family members will receive more activation due to spreading activation and, because of this spreading activation, they will more quickly reach their threshold activation level, resulting in a faster recognition.

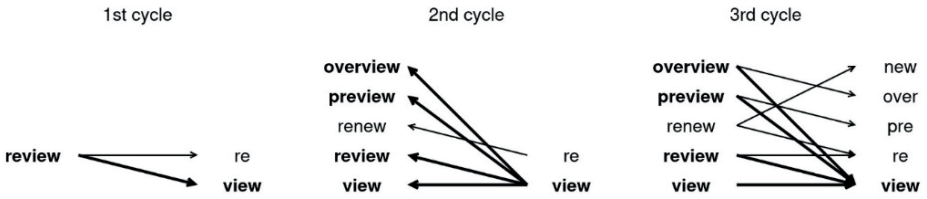


Figure 1.2 Schematic illustration of a general spreading activation mechanism. The presented word activates its root morpheme (1st cycle), which then activates the presented word's family members (2nd cycle). Family members start activating the root morpheme in the 3rd cycle. In the 4th cycle (not depicted here), the root morpheme will start to activate the presented word.

Whether the MFRM can account for the family size effect in spoken word recognition has not yet been investigated and is also not self-evident: the MFRM takes as its input written words as wholes, but spoken words unfold over time. There are strong indications that spoken words are incrementally processed as they become perceptible (e.g., Wurm, 1997; Kemps et al., 2005; ten Bosch, Boves, & Ernestus, 2022). To test whether a spreading activation mechanism between word and morpheme representations can account for the auditory family size effect, the MFRM would need to be adjusted in order to be applicable to an incrementally unfolding signal.

Most models of human auditory word recognition do not assume morpheme representations (e.g., Shortlist B, Norris & McQueen, 2008; Fine-Tracker, Scharenborg, 2008; Earshot, Magnuson et al., 2018; DIANA, ten Bosch, Boves, & Ernestus, 2022). The absence of morpheme representations in auditory word recognition models does not reflect a conviction, but rather the fact that previous research addressed other research questions than which linguistic units are represented in the mental lexicon. For instance, an important objective of Fine-Tracker (Scharenborg, 2008) was to account for effects of fine-phonetic detail (e.g., prosodic information). As localist explanations for the root frequency and the family size effect in written word recognition heavily rely on morpheme representations, more research is

necessary on how morpheme representations could be included in spoken word recognition models.

1.2.2 Distributional–connectionist theories in light of frequency and family size effects

In distributional–connectionist models, frequency effects arise as a consequence of how these models are trained. During training, a model is provided with the features ('cues') of word forms and these words' meanings (cf. Figure 1.1). The model then uses the cues and the association weights between cues and meanings to predict a word's meaning. This predicted meaning is then compared with the provided actual meaning, yielding a prediction error. Using optimisation algorithms such a backpropagation (e.g., Rumelhart et al., 1986) or the Widrow–Hoff learning rule (Widrow & Hoff, 1960; see also Rescorla & Wagner, 1972), the prediction error is used to improve the association weights between cue and meaning representations so that the next time the word is presented, the predicted meaning is closer to the actual meaning.

Distributional–connectionist models predict the meaning of higher frequency words more accurately than that of less frequent words, because higher frequency words are encountered more often during training, causing association weights to be updated more frequently and thus learned better. Accordingly, both in the recognition of written words (Seidenberg & McClelland, 1989; Heitmeier et al., 2024) and of spoken words (e.g., Gaskell & Marslen-Wilson, 1997; Shafaei-Bajestan, 2023), the accuracy with which the meaning of a word is predicted correlates with the frequency of that word, which is usually interpreted as an indication that distributional–connectionist models can explain the surface form frequency effect.

Following the same logic, distributional–connectionist models are also able to account for the root frequency effect. In these models, the association weights between sublexical cues and meanings are updated more often for roots that occur more frequently. As a result, when the word *catlike* is later presented, its cues can predict its meaning more accurately because the model has already built-up strong connections between the cues shared with the root and the related meaning of the root. Similar to hybrid localist models, distributional–connectionist models therefore predict that recognition is governed by both roots and surface forms, with their relative contribution depending on how strongly training has reinforced the associations linked to each.

That roots help to recognise morphologically family members enables distributional–connectionist models to explain the family size effect. The effect emerges because morphological family members provide a consistent learning environment for the mapping of a word's form cues onto the word's meaning. For illustration, all family members of *review* (e.g., *preview*, *viewing*) contain the form cue *view,* both in the visual domain and in the auditory domain. That is, all family members consist of either the letter trigrams *vie* and *iew* or the phoneme sequence /vju:/ or an audio sequence representing the word *view.* All family members also consist of a meaning that is closely related to the meaning of *view* **'to look at (something)'**. That is, *preview* refers to the act of **looking at something** before it is (publicly) available and *viewing*, as a noun, refers to the act or occasion of **looking at something** and, as a present participle verb, to the process of **looking at something**. Thus, all family members' meanings revolve around the concept of visual perception. Consequently, when training a distributional–connectionist model, the form cues of the root morpheme *view* tend to co-occur with the abstract meaning of 'to look at something', so that association weights can be optimised to represent the relationship between this form cue and this meaning. As a consequence, morphemes help to structure the mapping of word forms onto corresponding meanings.

Mulder et al. (2014) implemented an early version of the distributional–connectionist DLM (e.g., Chuang & Baayen, 2021) to predict visual lexical decision RTs. They found that RTs that were generated by their model correlated with family size to a similar extent as the observed RTs, indicating that a distributional–connectionist model can account for the visual family size effect. Because they did not analyse whether the predicted RTs explain all variance in the observed RTs that is also explained by family size, it is unclear whether a distributional–connectionist model can account for the complete family size effect or only for parts of the family size effect. In addition, it is currently still unclear whether distributional–connectionist theories can account for the family size effect in spoken word recognition.

1.3 Research methods

The recognition of morphologically complex words has been studied with a wide variety of methods, including behavioural experiments (for an overview see Amenta & Crepaldi, 2021) such as lexical decision (e.g., Meunier & Longtin, 2007; Winther Balling & Baayen, 2008, 2012) and word naming

(e.g., Plag & Baayen, 2009), as well as neuroimaging experiments (for an overview see Leminen et al., 2019) such as EEG (e.g., Leminen & Clahsen, 2014; Leminen et al., 2011; Smolka et al., 2013), MEG (e.g., Leminen et al., 2011), and fMRI (e.g., Lehtonen et al., 2009). In addition, experimental research has been complemented by computational modelling studies that test explicitly formalised assumptions about the word recognition process against human experimental data (e.g., Baayen et al., 2011; Mulder et al., 2014; Plag & Baayen, 2009). All of these research methods have their advantages and disadvantages, the discussion of which goes beyond the scope of this thesis. Instead, in what follows, I will discuss the methods used in this thesis and why I chose these methods.

All experimental data that is analysed in this thesis originate from lexical decision experiments. Lexical decision experiments aim to measure how long it takes for a participant to recognise a given word. To this end, participants are presented with stimuli that are either real words or pseudowords and are asked to decide as quickly and accurately as possible whether the stimulus is a real word of the language. As a response, participants press a button (*yes* or *no*), and the participants' RTs, defined as the latency between stimulus presentation and button press, are measured as well as whether the response is correct or not. It is assumed that longer RTs indicate that a word is more difficult to process.

Pseudowords play a crucial role in the lexical decision paradigm. They are needed to make two types of responses possible (real word or not). In addition, their properties determine how participants make their decisions. Suppose that all pseudowords only consisted of consonants, participants could easily discriminate between real and pseudowords on the basis of the presence versus absence of vowels. Therefore, pseudowords typically comply with the phonotactic and orthographic regularities of the language, resembling real words of the language, without being real words. This ensures that participants cannot base their responses on superficial cues alone and have to carefully process the stimulus. In many lexical decision experiments, a pseudoword may differ from a real word only in its last letter or sound. If so, participants learn, during the experiment, that this is the case. For auditory lexical decision experiments, this implies that participants learn to wait until the end of the word before making their decision, while in natural listening condition, listeners may recognise a word before its end (e.g., listeners may already recognise English *cathedral* after they have heard *cathed*).

Even though the lexical decision paradigm is frequently used in research on spoken word recognition, the interpretation of the results is not straightforward. The reason is that participants may make a *real word* response because a) they think that the stimulus is a real word but they do not know with certainty because they do not know the word, b) they know that the stimulus is a real word, or c) they know that the stimulus is a real word and they accessed the meaning of this word. While it can be assumed that option a) is unlikely when the participants are adult native speakers of the language, as in this thesis. The difference between b) and c) is of relevance and has to be kept in mind when interpreting lexical decision data. For instance, lexical decision RTs reflect that words are recognised the more quickly, the higher their frequency is (e.g., Brysbaert, Mandera, & Keuleers, 2018), but it remains unclear whether this only implies that listeners can more quickly determine for a high frequency than for a low frequency word that it is part of their vocabularies or also that they are faster in accessing the meanings of high frequency words. That is, it is unclear which cognitive processes (mechanisms) underly this frequency effect. Also, words usually have multiple meanings or nuances of a meaning and it is usually unknown which meaning is retrieved in a lexical decision experiment (e.g., the most frequent meaning or multiple meanings). For instance, the word *saw* can refer to a cutting tool, to the process of cutting something with this tool, to a short sentence that states something that is generally thought to be true, or to the act of seeing (i.e., as simple past form of *see*).

Despite its drawbacks, the lexical decision paradigm is still useful because it enables researchers to study the recognition of isolated words. In addition to this, the paradigm has some practical advantages for this thesis. First, lexical decision experiments can be conducted for written (e.g., Keuleers, Diependaele, & Brysbaert, 2010) and spoken word recognition (e.g., Ernestus & Cutler, 2015), because of which findings can directly be compared between modalities. Second, there is a rich body of work on the recognition of morphologically complex words in lexical decision experiments (e.g., de Jong, 2002; Clahsen, Sonnenstuhl, & Blevins, 2003; Meunier & Longtin, 2007; Milin et al., 2017). Studying complex word recognition with lexical decision experiments will thus allow for the comparison of my findings with previous findings. Third, the data of various large lexical decision experiments are publicly available for multiple languages such as Dutch (Ernestus & Cutler, 2015), English (Tucker et al., 2019), and French (Ferrand et al., 2018), which is why the resource intensive procedure of collecting data can be omitted.

In this thesis, I will analyse lexical decision RTs with either statistical or computational models. Statistical models allow for testing assumptions about the spoken word recognition process. For instance, predicting lexical decision RTs with a statistical model in the form of a regression that includes the predictor root frequency allows for testing whether words' root frequencies affect word recognition times: if root frequency explains a significant amount of variation in the RTs, this indicates that root frequency influences word recognition times and thus that root frequency plays a role in the recognition process.

Computational models are formalised and explicit descriptions of cognitive processes. Computational modelling here thus refers to the formal specification of representations and processes, independent of whether a model is implemented on a computer. Computational modelling, as understood in this thesis, is therefore distinct from computational linguistics in the narrow sense of engineering-oriented natural language processing, even though the two may share tools and techniques.

In this thesis, using the method of computational modelling, I will compare different word recognition theories by implementing these theories as computational models that specify how information is represented and transformed over time. More precisely, I will compare how accurately these models predict the lexical decision RTs from the same experiment. Computational modelling comes with the benefit that it brings to light when word recognition theories are underspecified. Such underspecifications become apparent because the implementation of a theory requires this theory to be unambiguous and explicit.

1.4 This thesis

This thesis will provide insights into the role of morphology in auditory word recognition. This thesis research questions are: First, does the morphological structure of a spoken word influence how listeners recognise this word? Second, which mechanisms underly effects of morphological structure in spoken word recognition? Third, which insights can be gained by investigating the recognition of spoken morphologically complex words with the research method of computational modelling? The insights obtained will inform research on and the further development of new and existing theories and models of human auditory word recognition.

As discussed in Chapter 1.1.1, there are indications that both the root and the surface form frequencies of morphologically complex words play a role in the recognition of these words. It is unclear, though, how root and surface frequency effects interact with each other in the word recognition process and which variables and mechanisms shape this interaction. Because of this, in Chapter 2, I will investigate three systematically different theoretical models for both written and spoken word recognition. The first model assumes that word recognition is always root-driven, the second that it is always surface form-driven, and the third that the ratio between root and surface form frequency determines whether recognition is root- or form-driven, with a clear tipping point marking the shift between the two. I will compare how well these models fit with both visual and auditory lexical decision RTs, gaining insights into which of the three assumptions represented by the three different models is the most plausible.

As discussed in Chapter 1.1.2, a plethora of studies suggests that words with more family members are responded to more quickly in written word recognition. In contrast, in spoken word recognition, the previous literature provides inconclusive findings with respect to the question of whether and how morphological family members affect the recognition of spoken words. Because of this, in Chapter 3, I will investigate whether the morphological family size effect also occurs in spoken word recognition. Previous research has investigated how orthographic and semantic similarities between the presented word and its family members affect the family size effect in written word recognition, suggesting that especially semantic similarities drive the visual family size effect. I will investigate whether this is also the case for the auditory family size effect, providing insights into the mechanisms that underly the auditory family size effect.

Because Chapter 3 will provide indications that family size affects the recognition times of spoken words, the question arises of how this effect can be explained. The previous literature provides two theoretical explanations for family size effects. One explanation is based on a distributional–connectionist viewpoint, assuming that morphemes structure the mapping of word forms onto corresponding meaning. This explanation has been implemented in the Discriminative Lexicon Model (Chapter 1.2.2) for written word recognition. In Chapter 4, I will investigate whether an implementation of the Discriminative Lexicon Model for spoken word recognition can account for the auditory family size effect.

The other explanation for the family size effect is based on a spreading activation mechanism, assuming that complex words propagate activation to associated words and vice versa. This explanation has been implemented in the localist Morphological Family Resonance Model (Chapter 1.2.1) for written word recognition. Because the Morphological Family Resonance Model cannot take an incrementally unfolding audio signal as its input, it cannot readily be applied to model auditory lexical decision RTs. Because of this, I will integrate a similar spreading activation mechanism into the localist, human spoken word recognition model DIANA (ten Bosch, Boves, & Ernestus, 2022) and investigate, whether the augmented DIANA can account for the auditory family size effect. Although the Discriminative Lexicon Model and the Morphological Family Resonance Model constitute systematically different explanations for the family size effect, these explanations have never been compared with each other on the basis of concrete model comparisons. Chapter 5 aims to close this knowledge gap by comparing the above-mentioned implementations of the distributional–connectionist Discriminative Lexicon Model (Chapter 4) and a localist spreading activation mechanism.

In Chapter 6, after discussing the experimental work I have carried out as part of this thesis in Chapters 2–5, I will conclude by discussing the insights I have gained in these chapters. In doing so, I will relate my research findings back to the question of the role of morphology in spoken word recognition (models). I will discuss the difficulties I encountered in conducting the experiments and derive questions for future research.

Chapter 2

A Tipping Point in Word Recognition? Investigating the Relationship Between Root and Form Frequency Across Visual and Auditory Modalities

This chapter is based on:
Müller, H., ten Bosch, L., & Ernestus, M. (2026). A tipping point in word recognition? Investigating the relationship between root and form frequency across visual and auditory modalities. *Morphology, 36*(1). https://doi.org/10.1007/s11525-025-09449-y

Abstract

For various theories of human word recognition, the question of how the recognition of morphologically complex words is influenced by the morphological root or the surface form of the word is of considerable relevance. According to many theories, (e.g., Baayen, Dijkstra, & Schreuder, 1997), the morphological root predominantly guides the recognition process unless the word is of a (relatively) high frequency of occurrence. We tested this 'tipping point' hypothesis by comparing a statistical model based on this hypothesis with two alternative statistical models: one assuming that word recognition is always root-driven and another assuming it is always form-driven. To this end, we modelled response time distributions from two large-scale lexical decision experiments in Dutch – one visual and one auditory – focusing on three suffixes: the plural suffix -en for nouns, the derivational suffix -heid for nominalisations, and -t as the second/third person singular present tense suffix for verbs. Our results indicate that words with the suffixes -t and -heid are retrieved as whole forms in both visual and auditory word recognition. In contrast, words with the suffix -en are best accounted for by both the root-driven and the form-driven models in auditory word recognition, while in visual word recognition, they support the tipping point hypothesis. Taken together, our findings suggest that both root-driven and form-driven principles are relevant for word recognition, while the assumption of a categorical tipping point is less tenable. This chapter contributes to our understanding of word recognition mechanisms in both localist and distributional–connectionist theoretical frameworks.

In normal and healthy conditions, we effortlessly process spoken words and, once we have learned to read, written words as well. Yet the mechanisms underlying word recognition are far from trivial and have been the subject of numerous studies for several decades (e.g., Morton, 1969; Grainger & Jacobs, 1996; Chuang & Baayen, 2021). A particular focus of research on word recognition has been on how we recognise morphologically complex words (for overviews see Amenta & Crepaldi, 2012; Milin, Smolka, & Feldman, 2018; Stevens & Plaut, 2022), that is, words that can be divided into smaller meaningful units (i.e., morphemes). For example, *walked* is complex because it can be broken up into the root *walk* and the suffix *-ed*. The present chapter contributes to research on how humans recognise morphologically complex words by investigating the roles of the frequencies of both the complex words (e.g., *walked*) and their roots (e.g., *walk*).

There are strong indications that the recognition of a morphologically complex word (henceforth: complex word) is affected by the word's morphological structure and morphologically related words. For instance, words (e.g., *walk*) are recognised the faster the more morphologically related words (e.g., *catwalk*, *walked, walkie-talkie, walkman*) exist in the language, an effect known as the morphological family size effect (e.g., Schreuder & Baayen, 1997; de Jong, 2002; Perdijk et al., 2012). To give another example, the short presentation of the root of a morphologically complex word before the presentation of the complex word (e.g., *walk – walked*) leads to a faster recognition of the complex word than when a morphologically unrelated word is presented beforehand (e.g, *talk – walked*; Geary & Ussishkin, 2018; De Grauw, Lemhöfer, Schriefers, 2019). Vice versa, the short presentation of a complex word before the presentation of this word's root (e.g., *walked – walk*) leads to a faster recognition of this root compared to when the root is preceded by a morphologically unrelated word (e.g., *talked – walk*; Beyersmann et al., 2016; Wilder et al., 2019). Importantly, such priming effects are significantly stronger for primes that are morphologically related to the target (e.g., *walked – walk*) compared to primes that are solely orthographically related to the target (e.g., *department – depart*), which indicates that priming is driven by the morphological but not the orthographic relationship between prime and target (e.g., Rastle et al., 2000; Cho, Pires, & Brennan, 2024).

This chapter focusses on the cumulative root frequency (henceforth root frequency), another measure based on morphological structure. The root frequency is the total frequency of all words that contain the presented word's root. Words are recognised more quickly when their root frequencies are higher

(henceforth r*oot frequency effect*, e.g., Taft, 1979; Meunier, & Segui, 1999; Solomyak & Marantz, 2010; Sánchez-Gutiérrez et al., 2018). Traditionally, it has been assumed that the root frequency effect occurs independently of the often-replicated effect of form frequency (e.g., Taft, 1979; Caramazza, Laudanna, & Romani, 1988; Colé, Beauvillain, & Segui, 1989; Alegre & Gordon, 1999), which is based on the number of occurrences of the word's surface form (e.g., Brysbaert, Mandera, & Keuleers, 2018). The assumption of independent root and form frequency effects is challenged, though, by studies suggesting an interaction between the two effects (e.g., Baayen, Wurm, & Aycock, 2007; Luke & Chistianson, 2011). For instance, Baayen et al. (2007) reported that, for English written words, higher root frequencies result in slower recognition times in high form frequency words and in quicker recognition times in low form frequency words.

2.1 Theoretical explanations for the root frequency effect (in interaction with form frequency)

Theories of how humans recognise morphologically complex words can be broadly categorised based on their view of the mental lexicon: either a localist view or a distributional–connectionist view. The localist view (e.g., Morton, 1969, 1970; Jackendoff & Audring, 2020; Taft, 2023) assumes that different linguistic units (e.g., morphemes, lemmas, words) have their own representations in the mental lexicon. The root frequency effect is the result of the involvement of the word's root in the recognition of the word, either because regular morphologically complex words are not stored themselves in the mental lexicon (e.g., Taft & Forster, 1975), or because morphologically complex words can be recognised both on the basis of their root and their own lexical representations (e.g., Baayen, Dijkstra, & Schreuder, 1997; Jackendoff & Audring, 2020; Taft, 2023), or because the lexical representations of complex words refer to the words' root (e.g., Meunier & Segui, 1999). The interaction of root frequency and form frequency can well be accounted for by theories that assume that both a word's own lexical representation and that of its root can be involved in word recognition. A word's root may play a more important role when the representation of the word itself is difficult to access, for instance because of its low frequency of occurrence (Baayen et al., 2007). In contrast, when the representation of the word is easy to access, the representation of the root may hinder word recognition because it activates morphologically related words as competitors for the word to be recognised.

According to the distributional–connectionist view (e.g., Seidenberg & McClelland, 1989; Gonnermann, Seidenberg, & Andersen, 2007; Blevins, 2016), the recognition of words does not involve subunit representations, and some distributional–connectionist models do not even assume lexical representations for words (Chuang & Baayen, 2021). Instead, word recognition is conceptualised as the mapping of an incoming signal (e.g., a letter sequence or an audio signal) onto a meaning representation. This mapping is often implemented in artificial neural networks that are trained with pairs of word and corresponding meaning representations. During training, given an input word representation, the network predicts a meaning representation, compares this prediction with the actual meaning representation, and then uses the difference between predicted and observed meaning to adjust itself so that it yields more accurate predictions when it encounters the same input the next time. In such a network, the root frequency effect emerges because root morphemes tend to have consistent form–meaning relationships. For instance, the words *catwalk*, *walked,* and *walkman* all contain the sequence *walk* and a meaning that is related to 'moving around'. In distributional–connectionist models, it is assumed that the most influential weights for the recognition of high frequency words are fine-tuned to these words themselves, whereas the most influential weights for the recognition of low frequency words are fine-tuned to these words' roots (i.e., more often encountered items are learned better; Seidenberg & McClelland, 1989). In order to be able to further develop both the localist and the distributional–connectionist theories, more insight is necessary in how root frequency affects word recognition, as a function of form frequency. Differences may be expected between visual and auditory word recognition since they differ from each other in crucial aspects.

2.2 Visual versus auditory word recognition

As mentioned above, morphological structure affects word recognition both in the written and the auditory modality. Nevertheless, we may expect differences between the two modalities. The fact that most written words can be perceived in their entire forms in one glance, while spoken words unfold over time, has two implications.

First, a written word's root can often be perceived immediately upon stimulus presentation, whereas a spoken word's root is always perceived after its prefix and before its suffix. As a consequence, a word's root may play a larger role

in the recognition of suffixed words than of prefixed words in the auditory modality, which is supported by Chapter 3 in the context of family size effects.

Second, whereas in the visual modality, the final segment of the word can be immediately perceived, in the auditory modality, listeners have to hear the complete word before they can hear the final segment. This affects how participants react in the many lexical decision experiments that have been conducted to investigate the role of a word's morphological structure, and in which the last segment of a stimulus can turn the stimulus from a real word into a pseudoword. Listeners tend to wait until word offset to make their lexicality decision (e.g., Ernestus & Cutler, 2015) and word duration is the strongest predictor of auditory lexical decision response times (RTs). In contrast, the most important predictor for visual lexical decision RTs is not word length in number of characters – the visual counterpart to word duration – but form frequency (e.g., Ferrand et al., 2018).

Another difference between written and spoken words is that, in many languages, the spelling of a root (e.g., *walk*) is typically independent of whether it is embedded in a morphologically complex word (e.g., *walked*). Nevertheless, the pronunciation of a root may depend on whether it is embedded in a complex word and in which word. To mention just a small possible difference, in spoken words of languages like English and Dutch, roots that are realised in isolation are longer than roots realised in complex words (e.g., Baayen et al., 2003). A more pronounced difference can be found in the vowel (of the root) in the word pair *breath* /**brɛθ**/ versus *breathing* /**bri**:**ð**ɪŋ/. These differences may make it more difficult to recognise a given root in different words in auditory word recognition, which may decrease the role of the frequency of the root in auditory word recognition.

How much the pronunciation of a root in a word differs from its most common pronunciation may vary with the affixes in the word. For instance, in Dutch, root-final obstruents may differ in their voicing depending on whether they are in syllable final position (together with a consonant initial affix) or in syllable-initial position, as a consequence of resyllabification with a following vowel initial affix.

2.3 Goal of the present chapter

The present chapter investigates the role of root frequency in the recognition of morphologically complex words, in both the written and the auditory modality, as a function of the suffix. Specifically, we compare the role of root frequency with the role of word form frequency. We do so in two ways. On the one hand, we compare a simple model of word recognition that only takes root frequency into account with a simple model that only takes word form frequency into account. On the other hand, we also test a simple model, the *meta model* (see Chapter 2.4), that assigns an important role to the relation between the two frequencies, assuming that root frequency is the only important predictor for words with relatively high root frequencies, and that word form frequency is the only important predictor for words with relatively high word form frequencies. We will refer below to this type of model as a model with a tipping point.

We analyse RTs from existing lexical decision data in Dutch. We chose Dutch because the meta model was already applied to Dutch (Baayen, Dijkstra, & Schreuder, 1997) and because of the availability of large datasets of both visual and auditory lexical decision data for this language. Similarly large datasets are also available for English (Keuleers et al., 2012; Tucker et al., 2019) and French (Ferrand et al., 2010; Ferrand et al., 2018), but we preferred Dutch, because it has a richer morphological system than English, and because it has a more transparent orthography–pronunciation relationship than French and English, which makes word recognition in the visual modality more similar to word recognition in the auditory modality.

We focus on words with three different suffixes (plural *-en*, verbal *-t*, and *-heid*), which differ in characteristics that may affect the hypothesised tipping point. This will shed light on how the characteristics of the affixes affect the word recognition process.

2.4 The tipping point model tested in the present chapter

We will test the role of the tipping point of root and form frequency on the basis of a computational model that was especially popular at the end of the last century, the meta model (Schreuder & Baayen, 1995; Baayen, Dijkstra,

& Schreuder, 1997). Despite its publication date of about 30 years ago, throughout the years, various authors have drawn on the meta model to explain their experimental findings on, among others, the processing of nouns in Italian (Baayen, Burani, & Schreuder, 1997), Hebrew (Vaknin-Nusbaum & Shimron, 2011; Vaknin-Nusbaum, 2025), Russian (Savinova & Malyutina, 2021), Dutch and German (Reifegerste, Meyer, & Zwitserlood, 2017), compounds in Finnish (Pollatsek & Hyönä, 2005) and Chinese (Zou et al., 2023), and English affixed free and bound roots (Solomyak & Marantz, 2009) as well as English suffixed words (Dawson, Rastle, Ricketts, 2018).

The meta model was developed to predict RTs to simplex (i.e., root words) and complex words such as plurals in Dutch and English. It assumes that recognition times consist of multiple time components. The first, constant, components are an *initial mapping time* for the mapping of the signal onto lexical representations (for simplex and complex words) and an *execution time* required to carry out a response. The third component is an *activation time*, the time it takes until a lexical representation reaches threshold activation level. For simplex words, the recognition of which is assumed to consistently be root-driven, the activation time is inversely proportional to the word's cumulative stem frequency. Baayen, Dijkstra, & Schreuder (1997) define cumulative stem frequency as the "summed frequencies of a stem and all words in which that stem occurs". A complex word's activation time would be inversely proportional to the word's form frequency in case of form-driven word recognition and to the word's cumulative stem frequency in case of root-driven word recognition. Above that, root-driven recognition would require additional processing time for the activation of the complex word's representation based on the activations of the root and morphologically related words. This time is called the *parsing penalty*. Note that such a parsing penalty would be equivalent to the process of deriving a complex word's meaning from similar word forms by means of analogy in distributional–connectionist models (e.g., Blevins, 2016).

The meta model assumes that root- and form-driven word recognition compete, and that the ratio of the two corresponding frequencies determines how complex words are recognised. When cumulative stem frequency is substantially higher than form frequency, the cumulative stem frequency effect is more likely to outweigh the parsing penalty and shift the recognition process toward root-driven processing. The meta model in its original form (Baayen, Dijkstra, & Schreuder, 1997) does not predict differences among words that only differ in their affixes. As mentioned above, however, root recognition may

be more difficult before certain suffixes than before others, which motivates our investigation of words with different suffixes.

2.5 Algebraic formulation of the tested models

Our statistical models – one with a tipping point, one only based on root-driven processing, and one only based on form-driven processing – follow Formulae (2.1), (2.2) and (2.3). They are grounded in the principles of mixed-effects modelling (e.g., Baayen, Davidson, & Bates, 2008) and predict RTs for morphologically simple and complex words using different principles. All formulae contain a by-participant random intercept (u_j) and make the RTs depend on characteristics of the word or the experiment that are known to affect word recognition times (e.g., word duration; $\beta_{1:n}x_{1:n}$, see below for which characteristics we implemented).

The predictors in addition to root and form frequency were not incorporated in the original 1997 formula of the meta model, and our formulation of the meta model thus deviates from the original formulation. We nevertheless incorporated them because they substantially improve the fit with the experimental data. Baayen, Dijkstra, and Schreuder (1997) meticulously controlled their stimuli for properties such as length in letters. We aimed to investigate the processing of a larger variety of stimuli, which better approximates processing of everyday language. We analysed RTs to stimuli that vary in length or duration, for instance.

Formula (2.1) shows the general specification of the root-driven model. It predicts each RT with an intercept (β_0) and with the root frequency ($\beta_{root}F_{root}$). In addition, the model assumes a parsing penalty (Δp), which covers the delay induced by recognising a complex word via its root (and morphologically related words). Formula (2.2) represents the general specification of the form-driven model. It differs from the root-driven model in two respects. First, it is the form frequency ($\beta_{form}F_{form}$), rather than the root frequency, that contributes to the prediction of RTs, and second, it does not have a parsing penalty. Formula (2.3) provides the general specification for the tipping point model. It combines (2.1) and (2.2) and states that the faster – root- or form-driven – recognition determines the predicted RTs. See Appendix A for more information. We based our frequencies on the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1996).

$$RT = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + u_j + \beta_{\text{root}} F_{\text{root}} + \Delta p \quad (2.1)$$

$$RT = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + u_j + \beta_{\text{form}} F_{\text{form}} \quad (2.2)$$

$$RT = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + u_j + min(\beta_{\text{form}} F_{\text{form}}, \beta_{\text{root}} F_{\text{root}} + \Delta p) \quad (2.3)$$

Unlike what may be suggested by Baayen, Dijkstra, and Schreuder (1997), the term *stem* is ambiguous in *cumulative stem frequency* because it is used in the literature to refer to inflectional bases or bound roots, among others (cf. Bauer, Lieber, & Plag, 2015). Because of this, in this chapter, we substitute cumulative stem frequency with root frequency. For computing root frequency, we sum the frequencies of the root and of all words with that root and maximally one other morpheme. For illustration, the root frequency of *spoons* is the sum of the frequencies of *spoon, spooned, spoonerism, spoonful, spoons,* and *tablespoon*, but not of the frequencies of *tablespoonful, cooking spoons*, or *spoon-feeding,* which consist of more than two morphemes. We did not include words with more than two morphemes in the root frequency computation because many words consisting of more than two morphemes have a frequency of zero or one in 42 million word tokens (Baayen, Piepenbrock, & Gulikers, 1996), and therefore are likely not stored in every individual listener's mental lexicon. Because of the low frequencies of words consisting of more than two morphemes, for our six datasets (see Chapter 2.6) the correlations between the root frequency based on words consisting of at most two morphemes and the root frequency based on all words ranges between 0.94 and 0.99.

In the modelling, we include the two most important characteristics of the words or experiment that predict RTs for lexical decision data. Our first control variable is the word's *duration in ms* for spoken words (Ernestus & Cutler, 2015) and the word's *length in number of letters* for written words (Keuleers, Diependaele, & Brysbaert, 2010), because it takes longer to respond to longer than to shorter words. Our second control variable is the *trial number* in each session (e.g., Ernestus & Cutler, 2015). This captures participants' gradual adaptation to the task throughout the experiment.

Appendix B shows that the three models as implemented with the formula (2.1), (2.2) or (2.3), respectively, can be correctly distinguished on the basis of a dataset of lexical decision times. It thus forms a proof a concept.

2.6 Outline of six studies

We present six studies, each focusing on a single suffix in either the visual or the auditory modality. Rather than analysing all data with a single statistical model with categorical predictors for suffix (*-en, -t, -heid*) and modality (*visual, auditory*), we deliberately analysed each dataset separately with independent statistical models. This decision was motivated by the fact that different theoretical processing models require different statistical formulations (see formulae 2.1–2.3). A unified model with categorical predictors would only identify the best-fitting model across all suffixes and modalities, masking variation in model performance across conditions. In contrast, separately analysing each dataset allows for testing, for example, whether the tipping point model best predicts RTs for one suffix in one modality, while a different model better predicts RTs for another suffix or modality. This approach enables us to test whether the presence of a tipping point depends on the suffix and/or the modality.

In Study 1, we studied written plural nouns that form their plural with the suffix *-en*. This suffix, which was also investigated by Baayen, Dijkstra, and Schreuder (1997), is very productive but can be easily confused with the infinitive suffix *-en,* the plural present tense verb suffix *-en*, and the inflectional suffix *-e* for adjectives, because these affixes are often homonymous. The suffix *-en* forms a syllable with the preceding consonants, which implies a difference in syllable structure between the root presented in isolation and the root in the inflected form. We expect that the tipping point model best predicts the RTs of the *-en* data, because this would be in line with Baayen, Dijkstra, and Schreuder's (1997) findings. In Study 2, we studied the same type of words in the auditory modality. We expect that the tipping point model best predicts the RTs again, because this would be in line with Baayen et al.'s (2003) findings.

In Studies 3 and 4, we studied second/third person singular present tense verb forms with the suffix *-t*, in the written and auditory modality, respectively. The suffix *-t* differs from the noun plural suffix *-en* (Studies 1 and 2) in several ways. First, while *-en* has multiple functions, *-t* predominantly marks the second/third person singular present tense. Second, while, phonologically, *-en* alters the syllable structure, *-t* does not. Finally, acoustically, *-t* may be more salient than *-en*, but both can be reduced.

Taken together, factors that may favour root-driven processing are present for both *-t* (no frequent homonyms with other morphological functions and, possibly, acoustic saliency) and *-en* (phonological saliency due to changes in the syllable structure). Because of this, differences between *-t* and *-en* are expected, but it is difficult to predict which of the two suffixes is more likely to induce root-driven recognition and therefore less likely to show evidence for a tipping point for a role of form frequency.

Finally, in Studies 5 and 6, we studied nouns that are formed with the suffix *-heid*, again in the visual (Study 5) and the auditory (Study 6) modality. The suffix *-heid* is the longest of all suffixes (both in terms of number of letters and phonemes); it does not have homonyms; it always forms a whole syllable on its own; it is always stressed; and it is the only derivational suffix among the suffixes under scrutiny and therefore it is the only suffix that adds an independent meaning to its base. Because of these differences, we expect words that are formed with *-heid* to have a greater likelihood to undergo root-driven processing (see also Baayen & Neijt, 1997) than words formed with *-t* or *-en* and therefore to be less likely to show evidence for a tipping point.

We based the studies on the visual modality on data from the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, 2010). DLP contains the responses and RTs from 39 native speakers of Dutch to 14,089 written Dutch content words and 14,089 written pseudowords. The words differ in word class and position of stress, among other characteristics. The pseudowords conform to the phonotactic rules of Dutch and are similar to the words in terms of word length and morphological structure.

We based the studies on the auditory modality on data from the Biggest Auditory Lexical Decision Experiment Yet (BALDEY; Ernestus & Cutler, 2015), representing an extensive auditory lexical decision experiment that was conducted in Dutch. BALDEY contains the responses and RTs from 20 native speakers of Dutch to 2,780 spoken Dutch content words and 2,761 pseudowords. The words systematically differ in word class, position of stress, number of syllables, and morphological structure. Each pseudoword was created by substituting one or two segments of a real word to make sure that the morphological and phonological structure was balanced across the word and pseudoword sets.

2.7 Methods

The data, the scripts used in this paper, and each model's implementation can be found at https://doi.org/10.34973/jm3a-vj10.

2.7.1 Data

We separately analysed RTs of subsets of DLP (Keuleers, Diependaele, & Brysbaert, 2010) and BALDEY (Ernestus & Cutler, 2015). These subsets were not controlled for any characteristics of the stimuli such as number of letters or duration in ms, or lemma, bigram, root, or form frequency.

To ensure that we restricted our studies to trials in which participants had processed the stimuli correctly, we only analysed RTs of correct responses that were not given earlier than 100 ms after stimulus onset in DLP and after stimulus offset in BALDEY. We chose these thresholds because the fastest human response time is assumed to be around 100 ms (Miller, 1968; Pain & Hibbs, 2007). In DLP, stimuli are presented at once and can thus be recognised as words on stimulus presentation (and the fastest possible RT is thus 100 ms after stimulus onset). In BALDEY, stimuli are incrementally presented due to their auditory nature, and the last speech sound segment could turn a word into a pseudoword. Thus, in BALDEY, stimuli can be recognised as words at its earliest at word offset (and the fastest possible RT is thus 100 ms after stimulus onset).

Table 2.1 *Number of stimulus types (and number of responses) per affix and per modality and the overlap in the two modalities.*

Affix	Modality	Simplex	Complex	Total
-en	Written	932 (30,706)	871 (26,579)	1,803 (57,285)
	Spoken	107 (1,804)	143 (2,547)	250 (4,351)
	Written & Spoken	59 (2,885)	86 (4,296)	145 (7,181)
-t	Written	1,140 (3,413)	222 (7,798)	1,362 (11,211)
	Spoken	99 (1,769)	33 (538)	132 (2,307)
	Written & Spoken	9 (469)	7 (391)	16 (860)
-heid	Written	187 (6,514)	55 (1,934)	242 (8,448)
	Spoken	48 (865)	80 (2,480)	128 (3,345)
	Written & Spoken	41 (2,179)	10 (683)	51 (2,862)

2.7.2 Plural nouns ending in *-en*

In Studies 1 and 2, in order to better estimate the effects of the control variables in the statistical model, we studied responses to both plural and singular nouns, although for investigating the tipping point hypothesis, studying plural nouns would be sufficient. The singular nouns in the visual and auditory datasets form their plural only with *-en* and the plurals end in *-en* and are not homophones of Dutch verbs. Due to Dutch regular spelling rules, the spelling of the root may differ between the singular and the plural: pluralisation may have resulted in the doubling of the singular's last consonant letter (*doubling*; e.g., *tak* 'branch' – *takken* 'branches') or removal of the singular's last vowel letter (*removal*; e.g., *boot* 'boat' – *boten* 'boats'). Furthermore, both datasets contain plurals that differ from the singular in the voicing of the stem-final obstruent (*devoicing*; due to final devoicing, e.g., *huis*/œys/'house' – *huizen* /œyzən/'houses').

The numbers of word types and the total numbers of responses are provided in Table 2.1 (first two rows). For Study 1 (DLP), Table 2.2 provides information about the root and form frequencies of the stimuli and about the relative frequencies for the plurals. More than half of the responses to plurals were elicited by plurals that do not just consist of the singular plus *-en*: 6,171 responses were given to plurals with *doubling* (23.22%), 4,905 responses to plurals with *removal* (18.45%), and 2,444 responses to plurals with *devoicing* (9.2%).

Table 2.2 *Descriptive statistics of the stimuli that were studied in Study 1. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ).*

DLP	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	51.70	28.08	54.64	10.58	0.366
SD	135.65	90.83	147.01	42.91	0.258
Min	0.05	0.02	0.02	0.02	0.001
Max	1,617.74	1,470.64	1,617.74	942.69	1.000

For Study 2, Table 2.3 provides the root and form frequencies of the stimuli and the relative frequencies for the plurals. Less than half of the plurals are not orthographically represented by simply the singular plus *-en*: 579 responses were given to plurals with *doubling* (22.73%), 687 responses to plurals with *removal* (26.97%), and 162 responses to plurals with *devoicing* (6.36%). Only the voicing is audible in these auditory stimuli.

Table 2.3 *Descriptive statistics of the stimuli that were studied in Study 2. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ).*

BALDEY	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	26.78	14.90	46.13	12.39	0.346
SD	86.97	77.64	77.64	25.94	0.228
Min	0.14	0.19	0.19	0.10	0.019
Max	574.50	407.17	407.17	181.86	0.955

Study 1 and 2 are to some extent similar to Baayen, Dijkstra, and Schreuder's (1997) study, testing the meta model too by investigating how well it can predict RTs. Our subset of DLP includes RTs to 91 of the 93 singular types and 92 of the 93 plural types tested by Baayen, Dijkstra, and Schreuder. These RTs correspond to 6,268 responses in our dataset. The subset of BALDEY includes RTs to only three of the singular types and ten of the plural types that were incorporated by Baayen, Dijkstra, and Schreuder (1997). Our datasets are much bigger because we also included plurals that not simply consist of the singular plus *-en,* but that also undergo *doubling*, *removal* or *devoicing*.

2.7.3 Second/third person singular present tense verb forms ending in *-t*

In Studies 3 and 4, we analysed responses to first person singular present tense verb forms, which just consist of a verb's root (e.g., *bak* 'to bake'; note that in a lexical decision experiment, these verb forms could also be interpreted as imperatives), and second/third person singular present tense verb forms, consisting of the root and *-t* (e.g., *bakt* 'bakes'). We excluded verbs with roots that end in *-t*, because their first, second, and third person singular present tense verb forms are identical. We also excluded the verbs *hebben* 'to have', *kunnen* 'can', *worden* 'to be (passive voice)', and *zullen* 'will', because they can function as auxiliary verbs and therefore have exceptionally high root frequencies. Table 2.1 shows the numbers of word types and total numbers of responses analysed in Study 3 and 4 (third and fourth row). The root and form frequencies of the stimuli and the relative frequencies for the plurals are provided in Table 2.4 for Study 3 and in Table 2.5 for Study 4.

Table 2.4 *Descriptive statistics of the stimuli that were analysed in Study 3. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ).*

DLP	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	140.02	8.02	151.77	15.33	0.200
SD	254.93	23.91	218.98	30.61	0.174
Min	0.26	0.02	0.52	0.02	0.005
Max	1,337.64	180.50	1,337.64	196.12	0.906

Table 2.5 *Descriptive statistics of the stimuli that were analysed in Study 4. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ).*

BALDEY	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	123.21	21.21	35.21	5.17	0.117
SD	371.48	74.51	84.01	16.79	0.131
Min	0.29	0.05	0.55	0.05	0.005
Max	2,484.93	559.88	477.69	91.62	0.605

2.7.4 Derived nouns ending in *-heid*

In Studies 5 and 6, we analysed responses to uninflected adjectives that can combine with *-heid* and nouns that consist of an adjective root and this suffix. Table 2.1 (fifth and last row) provides the numbers of word types and responses that were analysed. Table 2.6 provides frequency information about the stimuli analysed in Study 5 and Table 2.7 about the stimuli analysed in Study 6.

Table 2.6 *Descriptive statistics of the stimuli that were analysed in Study 5. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ).*

DLP	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	101.77	46.63	205.03	8.43	0.113
SD	223.06	114.86	363.72	18.46	0.139
Min	0.81	0.1	4.98	1.02	0.002
Max	1,835.81	1,091.10	1,835.81	120.71	0.675

Table 2.7 *Descriptive statistics of the stimuli that were studied in Study 6. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ).*

BALDEY	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	74.33	44.61	91.33	2.62	0.258
SD	185.76	119.31	258.52	4.59	0.358
Min	1.17	0.67	0.05	0.05	0.001
Max	1,206.36	751.40	1,835.81	33.14	1.000

2.8 Model fitting

All statistical models described in this paper were implemented in R (R Core Team, 2021). We implemented formulae (2.1), (2.2), and (2,3), representing the three models that we test, as Bayesian models, written in the language *stan* (Stan Development Team, 2022). We fitted the models with the package *cmdstanR* (e.g., Gabry & Češnovar, 2022) and its default Hamiltonian Monte Carlo (HMC) algorithm. We used four chains and as many samples as were needed to ensure that all values of R-hat were less than 1.01 and all effective sample sizes were above 400, which safeguards that the models' estimates are stable (using the criteria in Vehtari et al., 2021).

2.8.1 Model comparison

Model comparison was based on leave-one-out cross-validation using the expected log predictive density (ELPD LOO; Vehtari, Gelman, & Gabry, 2017). In Bayesian LOO, predictive accuracy is assessed by repeatedly considering each observation in turn as held out and evaluating how well the model predicts this observation using the posterior predictive distribution obtained from the remaining data. Importantly, these predictions marginalise over posterior uncertainty in the model parameters rather than relying on single point estimates. As a result, ELPD LOO reflects the expected predictive performance of the full Bayesian model, integrating over uncertainty induced by both finite data and model structure.

In practice, exact refitting of the model N times (once for each held-out observation) is computationally infeasible if N is large. Therefore, ELPD LOO was approximated using Pareto-smoothed importance sampling (PSIS-LOO; Vehtari, Gelman, & Gabry 2017), which estimates leave-one-out predictive accuracy directly from the posterior draws obtained via HMC. This approach

preserves the Bayesian nature of the evaluation: predictions are derived from the posterior predictive distribution and quantify how well the model generalises to unseen data, conditional on the full posterior rather than on maximum-likelihood estimates. This distinguishes Bayesian LOO from frequentist cross-validation procedures and from information criteria based on point estimates, such as AIC or BIC.

The higher the ELPD LOO, the better the model generalises to unseen data. We will graphically present how much the best performing model differs in ELPD LOO from the other models, accompanied by the standard error (SE) of these differences. Traditionally, these differences are expressed in the form of the ELPD LOO of a model minus the ELPD LOO of the best model (rather than the other way around; e.g., Gravel et al., 2024). Thus, differences in ELPD LOO are depicted as negative values.

Vehtari, Gelman, and Gabry (2017) favoured a model over another in case of an ELPD LOO difference of −10.2, which was twice as large as this difference's SE of 5.1, but they did not give any explanation for this threshold. We follow their criterion and assume that two models only systematically differ in their predictive performance when the absolute difference in ELPD LOO is at least twice the corresponding SE. In addition, because small differences in ELPD LOO are less informative (Sivula et al., 2020), we assume that the performance of two models can only be statistically distinguished if the difference in ELDP LOO is greater than −4 (cf. Vehtari 2020). Because the ELPD LOO does not penalise model complexity (Gronau & Wagenmakers, 2019), we prefer the less complex model when two models yield similar predictive performance according to our ELPD LOO criterion.

The data set of Study 1 is too large for the ELPD LOOs to be computed. In this case, we compute the ELPD LOO on the basis of a random subsample of the data that consists of one tenth of all observations, following Magnusson et al., (2020; see also Appendix B).

2.8.2 Prior estimation

Fitting Bayesian models requires the specification of prior distributions (henceforth *priors*) for the models' parameters (e.g., the effects of the frequency measures on the RTs and the size of the parsing penalty in the tipping point model and the root-driven model). These priors represent assumptions about possible and likely values of these parameters.

We determined the priors for the effects of length in number of letters or duration in ms, trial number, form frequency, and root frequency as well as by-participant random intercepts on the basis of the RTs of the correct responses to all morphologically simplex and complex stimuli in DLP and BALDEY, respectively, that closely resemble but are not identical to the stimuli to be studied in Studies 1–6. We analysed the RTs to these responses with Bayesian linear mixed-effects models, determining the coefficients of the predictors for which we wished to determine the priors. We chose uninformative priors for the predictors, that is, normal distributions with mean = 0, SD = 0.2, and a positive prior of *exp(0.1)* for the SD of the by-participant intercept. As in Studies 1–6, we excluded RTs that were smaller than 100 ms from word onset (DLP) or from word offset (BALDEY), and log-transformed all numerical variables.

The dataset from the DLP contains 160,906 responses with 116,320 responses to 3,627 unique simplex words and 44,586 responses to 1,436 unique suffixed words. Table 2.8 shows that word length in letters, root frequency, and form frequency yield facilitative effects, whereas the effect of trial number is inhibitory. The estimated coefficients shown in Table 2.4 were used as priors for all models reported below that were fitted to RTs in DLP.

Table 2.8 *Estimated effect sizes and standard deviations (SD) for the variables of interest in DLP. These estimates were used as priors in Experiments 1, 3, and 5.*

	Intercept	Length	Trial	Root frequency	Form frequency
Coefficient	6.5173	−0.0075	0.0075	−0.0042	−0.0213
Coefficient SD	0.0658	0.0026	0.0006	0.0003	0.0003

The dataset for determining the priors for BALDEY contains 7,556 responses, with 378 responses to 21 unique simplex words and 7,556 responses to 440 unique suffixed words. Table 2.9 shows that the estimated effects of duration in ms and trial number are inhibitory, while root and form frequency yield facilitative effects. The estimated effects shown in Table 2.9 were used as priors for all models reported below that were fitted to RTs of BALDEY.

Table 2.9 *Estimated effects and standard deviations (SD) for the variables of interest in BALDEY. These estimates were used as priors for models applied to BALDEY in Study 2, 4, 6.*

	Intercept	Duration	Trial	Root frequency	Form frequency
Coefficient	5.7090	0.2176	0.0127	−0.0054	−0.0091
Coefficient SD	0.1016	0.0125	0.0029	0.0018	0.0016

We tested different priors for the parsing penalty, because we had no knowledge about nor strong indications of what would be good priors for the parsing penalty and we wished to rule out that the priors for the parsing penalty affect the estimate of the parsing penalty, and consequently the goodness of the model's fit to the data. Assuming that recognising a complex word based on its root (and morphologically related words) takes between 1 ms and 350 ms, we tested 350 normal distributions as parsing penalty priors, each with a mean ranging between 1 ms and 350 ms and a standard deviation SD_{prior} as given in Equation (2.4). In Equation (2.4), both mean and SD are expressed in ms. Equation (2.4) ensures the standard deviation is approximately one fourth of the mean for larger means, but is never shorter than 1 ms. If SDs were smaller than 1 ms, the prior distributions would be too peaked so that the information in the data could not overrule the information provided by the prior. For more details about how we computed the priors, see Appendix C.

$$SD_{\text{prior}} = \frac{\text{mean}}{4} + \frac{4}{\text{mean} + 4} \tag{2.4}$$

2.9 Results

Figure 2.1 shows, for Studies 1–6, the difference in ELPD LOO between the root-driven model, the form-driven model, and the tipping point model relative to the model that predicts the RTs the most accurately (i.e., the model with the highest ELPD LOO). In Study 1 (written *-en* in DLP), the tipping point model predicts the RTs most accurately. In Study 2 (spoken *-en* in BALDEY), the root-driven and the form-driven model predict the RTs equally accurately; the tipping point model may yield the same predictive accuracy as the other two models, but it is more complex and thus considered to be a worse fit to the data.

In Studies 3–6 (written/spoken *-t*/*-heid* in DLP/BALDEY), the form-driven model predicts the RTs most accurately. The tipping point model is considered to fit the data worse than the form-driven model in Studies 3–6 because it draws on a greater complexity to yield the same prediction accuracy as the form-driven model. Moreover, the tipping point model actually behaves like the form-driven model because the parsing penalty is so high that the recognition of complex words is always form-driven and never root-driven (see also Appendix B).

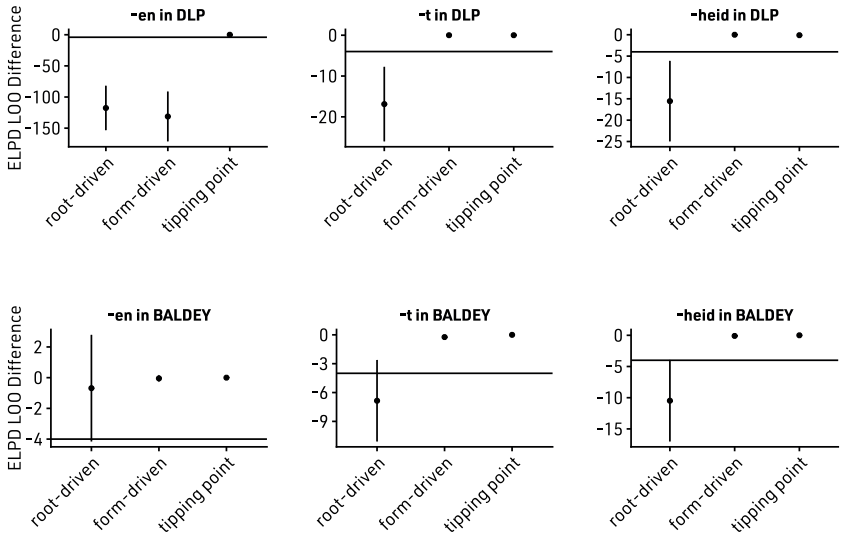


Figure 2.1 Difference in ELPD LOO (y-axis) between the three processing models (x-axis) and the model(s) with the highest ELPD LOO (which can be two models at the same time), for each suffix (columns) in the two modalities (rows: visual in top row, auditive in bottom row). Error bars represent two times the standard error. The solid horizontal line in each panel represents a difference in ELPD LOO of minus four, our threshold for determining whether two models significantly differ in how accurately they fit the data.

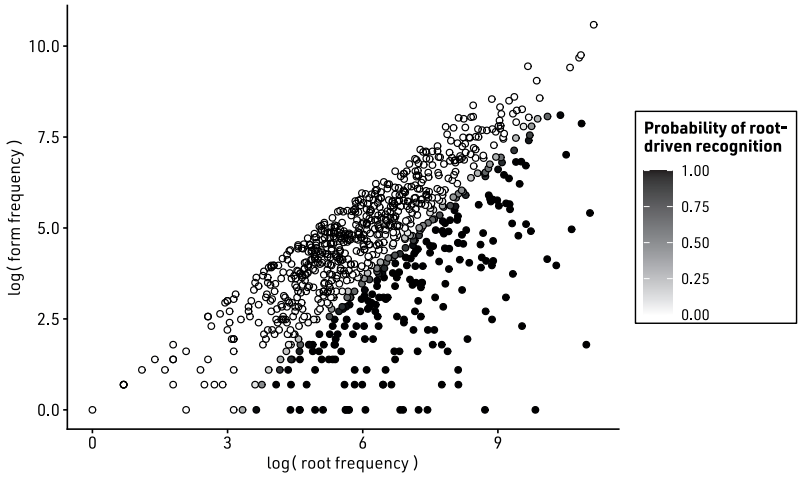


Figure 2.2 Probabilities of root-driven recognition (indicated by marker grey level) for written plural nouns ending in *-en* as a function of the plurals' log-transformed root frequencies (x-axis) and form frequencies (y-axis). The probabilities are based on the posterior distribution of the tipping point model. Clearly, the tipping point depends on both root frequency and form frequency.

The results from all studies are independent of the choice of the prior mean for the parsing penalty. We describe the relevant aspects of the results of the six studies in more detail in the following subsections. Summaries of all models fitted can be found in Appendix D. The coefficients and their p-values of the control variables are not further discussed here as they are not of interest for answering our research questions.

2.9.1 Study 1: written plural nouns ending in *-en*

As mentioned above (and as also shown in Figure 2.1), RTs to written nouns ending in *-en* are predicted the most accurately by the tipping point model. In this model, the estimated effects of both root frequency (mean = −0.0291, SD = 0.0005) and form frequency (mean = −0.0261, SD = 0.0006) are facilitative. The estimated parsing penalty for the tipping point model is 59 ms. With this estimated parsing penalty, the recognition of 14% of the plurals is always root-driven and the recognition of 28% of the plurals is always form-driven. For 58% of the plurals, both root-driven and form-driven recognition is possible.

As can be seen in Figure 2.2, the root-driven recognition only occurs when the complex word's log-transformed root frequency is at least 3.14 (i.e., 0.55 per million tokens). Crucially, whether recognition is root-driven or form-driven depends on the relative frequency (i.e., form frequency divided by root frequency). Relative frequencies lower than 0.101 result in obligatory root-driven recognition and relative frequencies higher than 0.337 result in obligatory form-driven recognition. Plural nouns with relative frequencies between these two values can be recognised with either processing mechanism.

2.9.2 Study 2: spoken plural nouns ending in *-en*

As mentioned above (and see also Figure 2.1), RTs to spoken nouns forming their plural with *-en* are equally accurately predicted by the root-driven model and the form-driven model (the tipping point model relies on a greater complexity than the other two models and is thus disregarded). The estimated effect of root frequency in the root-driven model (mean = −0.0063, SD = 0.0012) is facilitative, as is the estimated effect of form frequency in the form-driven model (mean = −0.0053, SD = 0.0014). For plurals, log-transformed root and form frequency strongly correlate with each other ($r(2545) = .864$, $p < .001$), and, therefore, the similar performance of the root-driven model and the form-driven model may be due to a high correlation between root and form frequency.

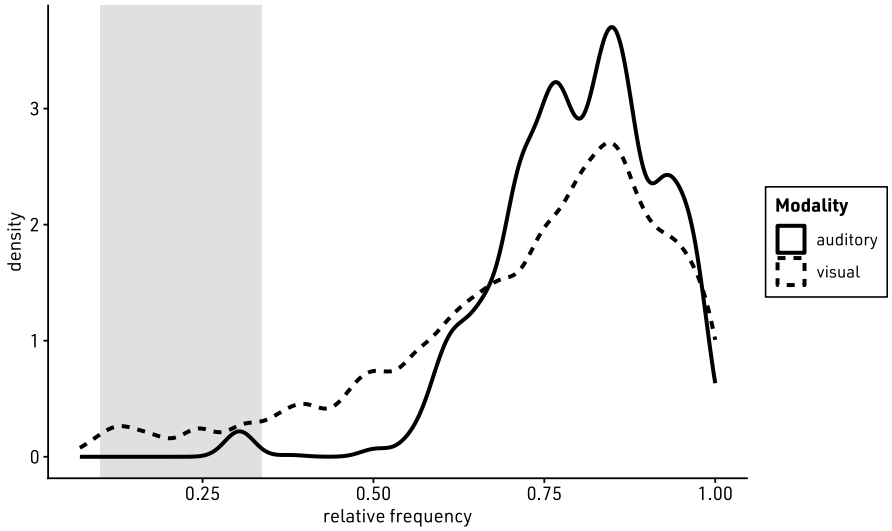


Figure 2.3 Plurals' relative frequencies (x-axis) and corresponding densities (y-axis) for DLP and BALDEY (line type). The grey-shaded area marks relative frequencies between 0.101 and 0.337, which were associated with flexible processing mechanisms (root- and form-driven) in Study 1.

Figure 2.3 shows the relative frequencies of the plural nouns in the written (Study 1) and spoken (Study 2) datasets. Compared to the visual dataset, the auditory dataset provides poorer coverage of verb forms with relative frequencies between 0.101 and 0.337 – forms that, as discussed in Chapter 2.9.1, can be processed either root- or form-driven. This may be the reason why the tipping point model is supported by the visual data but not by the auditory data.

2.9.3 Studies 3–6: written/spoken words ending in *-t/-heid*

As mentioned above (and shown in Figure 2.1), RTs to second/third person singular present tense verb forms ending in *-t* and RTs to derived nouns ending in *-heid* are the most accurately predicted by the form-driven model in both the visual and the auditory modality. The effect of form frequency is facilitative regardless of suffix and modality (see Tables D.1–D.6 in Appendix D). Correspondingly, Studies 3–6 do not provide support for the tipping point model.

Figure 2.4 Complex words' relative frequencies (x-axis) and corresponding densities (y-axis) for Studies 1, 3 and 5 (first plot; line type) and Studies 1, 4 and 6 (second plot; line type). The grey-shaded area marks relative frequencies between 0.101 and 0.337, which were associated with flexible processing mechanisms (root- and form-driven) in Study 1.

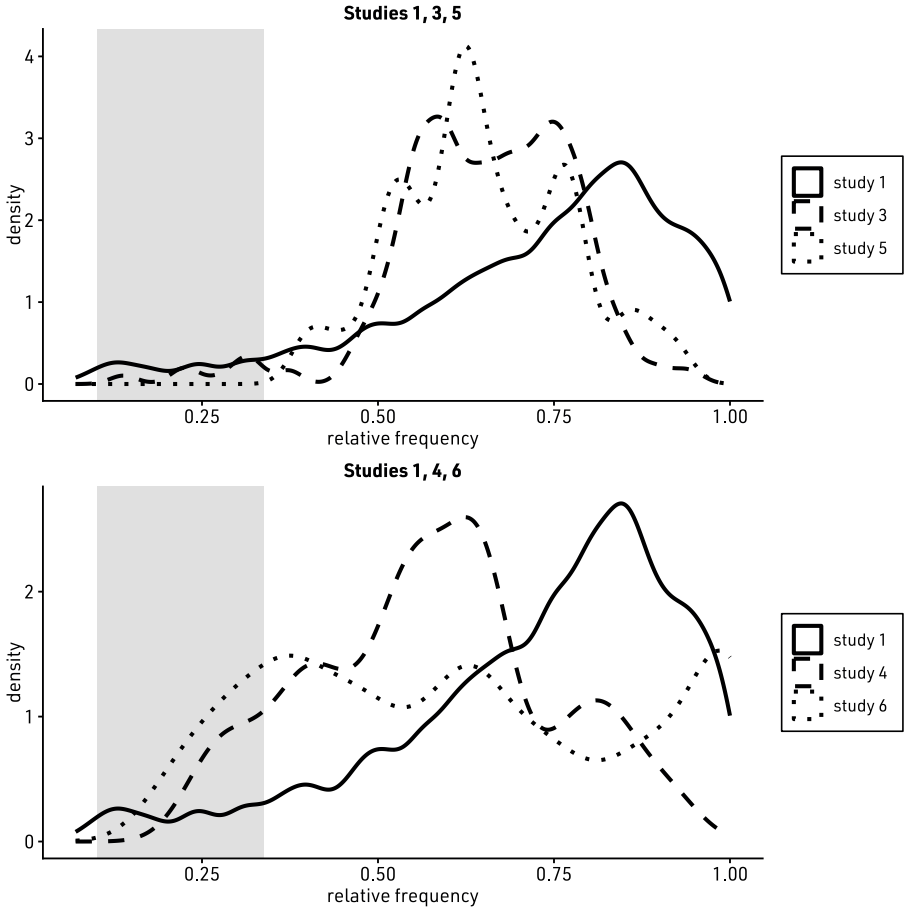


Figure 2.4 visualises the density distributions of the relative frequencies of the complex words examined in Studies 3–6, with Study 1 included as reference. It shows that stimuli with relative frequencies between 0.101 and 0.337 – those that can be processed via either their root or form (cf. Chapter 2.9.2) – make up a smaller proportion of the stimulus set in Studies 3 and 5 than in Study 1 (first plot in Figure 2.4). In contrast, this frequency range is more heavily represented in Studies 4 and 6 (second plot in Figure 2.4). Crucially, the fact that the tipping point model does not always account the best for the data, even when many stimuli fall within this range, suggests that relative frequency alone does not fully determine model fit.

2.10 General discussion

We investigated how root frequency and form frequency of three types of Dutch morphologically complex words affect how they are recognised. Specifically, we examined whether there is a tipping point in the relative frequency of root versus form that determines whether processing is predominantly root-driven or form-driven. To this end, we analysed lexical decision response times (RTs)

to both written and spoken words consisting of a root and one of three suffixes (*-en*, *-t*, *-heid*). These RTs were modelled using three different statistical approaches: one reflecting the tipping point hypothesis, and two others representing the assumptions that processing is either consistently root-driven or consistently form-driven.

We found that four of the six analysed datasets (written and spoken *-t* and *-heid*) are best fitted with a form-driven account. One dataset (spoken *-en*), for which form and root frequency are highly correlated, is equally well predicted by both the root-driven and the form-driven account. These results lend support to the idea that surface forms play a more important role in the word recognition process than morphological roots.

Interestingly, visually presented nouns ending in *-en* lend support to the tipping point hypothesis. Word recognition is consistently form-driven for plurals with root frequencies that are less than three times higher than their form frequencies, and root-driven for words with root frequencies that are more than ten times higher than the corresponding form frequencies. The recognition of visually presented plurals with root frequencies that are about three to ten times higher than their form frequencies (relative frequency between 0.101 and 0.337) can be both root- and form-driven. This finding suggests that both morphological roots and surface form may be of importance for the word recognition process.

Our finding that written Dutch plural nouns ending in *-en* support both root- and form-driven processing dovetails with the results reported by Baayen, Dijkstra, and Schreuder (1997). This raises the question of why plural nouns ending in *-en* lend support to both root-driven and form-driven processing, whereas second/third person singular present tense forms ending in the suffix *-t*, and nouns that are derived from adjectives with the derivational suffix *-heid* support form-driven processing in both the visual and the auditory modality.

2.10.1 The exceptional status of nouns ending in the plural suffix *-en*

Both the plurals ending in *-en* and the verb forms ending in *-t* are inflectional forms and/or semantically transparent (in contrast to *-heid*, which is derivational and semantically less transparent). This rules out that one of these properties is the reason that plurals are recognised differently. The affixes *-en* and *-heid* both lead to resyllabification of the word and both contain a vowel. So resyllabification and the presence of a vowel also do not seem to

account for the mode of processing. Baayen, Dijkstra, and Schreuder (1997) suggest that the lower a complex word's relative frequency is, that is, the lower the complex word's frequency is in comparison to its root frequency, the more likely the complex word undergoes decomposition (cf. also Hay 2001; Hay & Baayen, 2002). As can be seen in Table 2.3, 2.8, and 2.11, the *-en* stimuli have higher relative frequencies (mean = 0.366, SD = 0.258) compared to both the *-t* stimuli (mean = 0.2, SD = 0.174) and *-heid* stimuli (mean = 0.113, SD = 0.258). Our results therefore also do not support the hypothesis that a low relative frequency may lead to a role for root frequency in the recognition of at least some forms.

It is especially the plurals with relative frequencies between 0.101 and 0.337 that show effects of both root and form frequencies. Plurals with lower relative frequencies mostly show root frequency effects and plurals with higher relative frequencies mostly form frequency effects. The visual plurals in the relative frequency range from 0.101 to 0.337 do not seem to have other properties than the other plurals, except for relative frequency. The vast majority of the spoken plurals and the *-t* and *-heid* words presented visually have relative frequencies that are (much) higher than 0.337. This could explain why these words do not show root frequency effects. The orally presented *-t* and *-heid*, in contrast, are represented with many words with frequencies below 0.337. Possibly, for spoken words, root frequency plays a less important role than for written words, because the root is less easily to segment from spoken words, as a result of the acoustic differences between roots spoken in isolation and in different complex words (including differences in duration and co-articulation) that are not visible in the spellings of these Dutch words.

Another explanation for why written *-en* favours the tipping point hypothesis in contrast to *-t* and *-heid* is that *-en* is more productive than *-t* and *-heid*. It does not only form plurals for nouns, but also for verbs, and it can be used to derive nouns from verbs. A relationship between productivity and mode of processing forms another explanation (in addition to the high correlation between root and form frequency) of the finding that the form-driven and the root-driven model predict RTs to spoken words ending in *-en* equally accurately. There are indications that affix productivity increases the likelihood of root-driven processing in Dutch and English (Bertram, Schreuder, & Baayen, 2000; Bertram, Lalne, & Karvinen, 1999; Hay & Baayen, 2002). In Arabic, the productivity of root templates appears to facilitate root-driven processing

as well (Boudelaa & Marslen-Wilson, 2011), pointing again to a relationship between productivity and processing mode.

2.10.2 The precise formulation of the tipping point model

Our tipping point model, based on the meta model by Baayen, Dijkstra and Schreuder (1997), predicts that the roles of root and form frequencies mostly depend on the basis of the relative values of these frequencies. The question arises whether this formulation of the tipping point model is too simplistic to predict when root and form frequencies contribute to a word's recognition. Our results suggest that there is a difference between the visual and the auditory modality. In addition, there may be a role for the productivity of the affix. A new and more elaborated formulation of the tipping point model may take these factors into account.

Another issue with our implementation of the tipping point model concerns the exact definition of the root frequency. In the present chapter, we determined root frequency on the basis of words consisting of at most two morphemes. Including also words that contain more than two morphemes yields comparable frequencies as indicated by high correlation coefficients between the two frequency measures ($0.94 < r < 0.99$, depending on the analysed dataset), because of which it may not make a difference whether one includes words consisting of more than two morphemes or not. Such a difference may be observed, though, in the context of highly agglutinative languages, in which words regularly consist of more than two morphemes.

Although the tipping point model only performs better than a root-driven or a form-driven model for one of the analysed datasets, this is a clear indication of a trade-off between root-driven and form-driven processing. In light of our results for the visually presented plurals, which are sometimes recognised by their roots and sometimes by their forms, it is questionable though, whether such a trade-off is categorical in nature (i.e., recognition is either root-driven or form-driven), as assumed in our tipping point model. It could be that root-driven and form-driven processing are two poles of a continuum with mixed processing mechanisms in between.

2.11 Conclusion

In the present chapter, we tested for the first time on the same datasets whether a tipping point model better describes the processing of both written and spoken morphologically complex words than a root-driven and a form-driven model. We formulated the three theoretical processing models as statistical models following Baayen, Dijkstra, and Schreuder (1997) and identified the models that best fit the distributions of human lexical decision data for Dutch complex words. Our results suggest that the recognition of most complex words, in both visual and auditory word recognition, is form-driven, but root-driven processes seem to play a role too, at least in the processing of visually presented plural nouns ending in *-en*. We argue that relative frequency does not determine on its own whether recognition is root- or form-driven, but that it interacts with other factors such as the ease with which the root can be identified and affix productivity. Moreover, our results do not support a categorical tipping point that determines whether word recognition is either root-driven or form-driven, but rather a continuum between root- and form-driven recognition. Because distributional–connectionist theories do not posit a strict boundary between processing modes, whereas localist theories often do, accounting for such a continuum seems to be a bigger challenge for localist theories than distributional–connectionist theories of human word recognition

2

Chapter 3

The Family Size Effect in Visual and Auditory Word Recognition

This chapter is based on:
Müller, H., ten Bosch, L., & Ernestus, M. (2024). The family size effect in visual and auditory word recognition. *Language, Cognition and Neuroscience, 39*(6), 793–814. https://doi.org/10.1080/23273798.2024.2337941

Abstract

Words with larger morphological families elicit shorter response times in visual lexical decision (e.g., Baayen, Lieber, & Schreuder, 1997), but in auditory lexical decision, family size effects have been reported to be facilitative (Winther Balling & Baayen, 2008), absent (Baayen, Wurm, & Aycock, 2007), or inhibitory (Winther Balling & Baayen, 2012). We revisit the auditory family size effect and different family size definitions by re-analysing data from two large lexical decision experiments. The results suggest that words with larger families elicit smaller response latencies not only in visual but also in auditory lexical decision. However, in auditory lexical decision, this effect is insignificant when the word contains a prefix. In both modalities, it is the degree of semantic similarity and the degree of form overlap between a word and its family members, rather than the mere presence of overlap, that seem to drive the family size effect.

A word's family size is the type count of all words in which the given word's root occurs as a constituent. In many languages, family size can vary considerably by word. For instance, the Dutch word *praat* 'talk' has nearly 400 family members listed in CELEX (Baayen, Piepenbrock, & Gulikers, 1996), such as *praatgraag* 'talkative', *bijpraten* 'to catch up with somebody', or *gekkenpraat* 'fool's talk'. The noun *alinea* 'paragraph', in contrast, only appears in *slotalinea* 'final paragraph' and the corresponding plurals *alinea's* and *slotalinea's*. A word's family size is a good predictor for how quickly language users comprehend written words. For instance, in lexical decision experiments, participants respond faster to words with larger morphological families (e.g., Schreuder & Baayen, 1997; de Jong, Schreuder, & Baayen, 2000; Juhasz & Berkowitz, 2011). In contrast, little is known about whether and how a word's morphological family affects word processing when the word is auditorily rather than visually presented. Not only are there but few studies that addressed the effect of family size in auditory word recognition (Baayen, Wurm, & Aycock, 2007; Winther Balling & Baayen, 2008, 2012; Wurm et al., 2006), but they also provide conflicting results. It is not self-evident that a word's family size also affects that word's processing in auditory word recognition because spoken words unfold over time. As a consequence, a word's root, which is the most important cue to the morphological family, is not immediately perceptible.

This chapter aims to clarify the discrepancies between the results observed in the literature on the family size effect in auditory word recognition. In addition, this chapter aims to shed light on the mechanisms underlying the family size effect in both the visual and the auditory modality by testing different family size measures and analysing the family size effect for words with different morphological structures.

In the following, we will review the literature on the visual and auditory family size effect. We will then discuss the mechanisms underlying the family size effect. Based on the review and the discussion of the mechanisms, we will derive the research questions for the present chapter and elaborate on how these research questions will be answered.

3.1 Family size effect in visual versus auditory word recognition

The *visual* family size effect has been reported for languages with a concatenative morphology like Dutch (e.g., Moscoso del Prado Martín et al., 2004, 2005; Perdijk et al., 2012; Mulder, Dijkstra, & Baayen, 2015), English (e.g., Dijkstra et al., 2005; Kuperman et al., 2009; Baayen et al., 2011), and German (Lüdeling & de Jong, 2002); for Finnish, which has a particularly rich concatenative morphology (e.g., Moscoso del Prado Martín et al., 2004, Kuperman, Bertram, & Baayen, 2008); for languages with a non-concatenative morphology like Hebrew (Moscoso del Prado Martín et al., 2005) and Arabic (Boudelaa & Marslen-Wilson, 2011); and in the largely analytic language Mandarin (Feldman & Siok, 1997). The family size effect is elicited by words with different morphological structures like simplex words (e.g., Baayen, Feldman, & Schreuder, 2006; Baayen et al., 2011; Mulder, Schreuder, & Dijkstra, 2013), words that consist of a root and a suffix (e.g., Bertram, Baayen, & Schreuder, 2000; de Jong, Schreuder, & Baayen, 2000), or compounds (e.g., de Jong, Schreuder, & Baayen, 2000; Juhasz & Berkowitz, 2011), and when words that consist of a prefix and a root, a root and a suffix, and simplex words are investigated without further differentiation (Moscoso del Prado Martín et al., 2004). Larger families also facilitate the recognition of Dutch past participles, most of which are circumfixed (Bertram, Baayen, & Schreuder, 2000; de Jong, 2002).

The potential influence of family size on auditory processing remains uncertain, as the effects observed in the visual domain may not directly translate. In visual processing, due to parafoveal preview (e.g., Rayner, 1998), information about upcoming segments becomes available before the fixation of these segments. Consequently, the root can be quickly identified, even when it is preceded by a prefix. This could explain why the family size effect has been observed across various morphological structures. Conversely, in the auditory modality, it may take longer to identify a word's root because spoken words incrementally unfold and therefore roots will be identified after prefixes. Additionally, family members that show form overlap with the presented word may play a more important role. Specifically, prefixed words may primarily facilitate the recognition of other prefixed words, while words without prefixes predominantly influence the recognition of other words lacking prefixes. This hypothesis finds partial support in the four studies we encountered in the literature investigating family size effects in the auditory domain.

Baayen, Wurm, & Aycock's (2007) study set out to investigate differences between visual and auditory word recognition. They presented in a lexical decision experiment English written and spoken words made of a prefix and a root, or a root and an inflectional or a derivational suffix. Most of these words' family members did not contain a prefix. The visual experiment showed a facilitative family size effect, whereas there was no effect in the auditory experiment. Because the auditory and visual experiments presented the same stimuli, the absence of an auditory family size effect cannot be ascribed to the stimuli.

Wurm et al. (2006) conducted a study on the recognition of Dutch prefixed words. They found that family size facilitated word recognition for words whose uniqueness point was independent of whether the competitors only included words with the same prefix or all words starting with the same sounds. For the other prefixed words, the family size produced a null effect.

Two studies investigated lexical decision in Danish. Winther Balling & Baayen (2008) found that suffixed words (derived or inflected) exhibited a facilitative family size effect while simplex words did not. Winther Balling & Baayen (2012) conducted two lexical decision experiments. The first experiment investigated compounds and prefixed words. Family size was only found to facilitate responses when the family size was calculated on the basis of those family members that contain the shared constituent in word-initial position. This subset of the morphological family, known as onset-aligned family, supports the hypothesis that in auditory word recognition, the family size effect is mostly driven by members that show form overlap with the presented word from onset onwards. Experiment 2 investigated prefixed and suffixed words and produced an inhibitory family size effect. At least for suffixed words, the inhibitory effect was driven by those family members in which the shared constituent was not in word-initial position. This family size was strongly correlated with the classical family size ($r = .95$).

All in all, the results discussed in this chapter suggest that the family size effect varies between visual and auditory word recognition. In visual word recognition, the family size effect is always facilitative and largely independent of the word's morphological structure. In auditory word recognition, in contrast, the family size effect varies across words with different morphological structures, and whether or not the common constituent of the family members occurs in word-initial position seems to play a role too.

3.2 Mechanisms underlying the family size effect

Several researchers have formulated explanations for the family size effect, which they implemented with computational models. These models successfully simulate the visual family size effect. We are aware of two conceptually different computational models: the *Morphological Family Resonance Model* (MFRM; De Jong, Schreuder, & Baayen, 2003) and the *Naïve Discriminative Reader* (NDR; Mulder et al., 2014). The latter is very similar to *Naïve Discriminative Learning* (NDL; e.g., Baayen et al., 2011), which has been developed into *Linear Discriminative Learning* (LDL; e.g., Baayen et al., 2019).

The MFRM operates within the framework of spreading activation. In the MRFM, a word's lemma representation (cf. Levelt, 1989) is connected to syntactic and semantic representations (cf. Schreuder & Baayen, 1995). An activated lemma representation propagates activation to connected syntactic and semantic representations, which in turn propagate activation to connected lemma representations, including the original lemma representation. Once a word's lemma representation reaches threshold activation level, this word is recognised. A lemma representation receives activation from more lemma representations via shared semantic representations, the larger the word's morphological family is. Consequently, lemma representations for words with larger morphological families reach the threshold activation level earlier than lemma representations of comparable words with fewer family members, resulting in a faster recognition and a faster response in lexical decision experiments.

One characteristic of the MFRM is that it assumes that all family members equally contribute to the family size effect. The literature, however, suggests that this is not the case. For example, Moscoso del Prado Martín et al. (2004) suggest that only semantically related family members contribute to the family size effect. They analysed visual lexical decision data from Finnish and distinguished two different family size measures. One measure is only based on the family members that have a semantic relationship with the presented word, while the other is based on the remaining family members. The results show that only the first measure significantly predicted response latencies. Similar findings were obtained for Dutch (Schreuder & Baayen, 1997; Bertram, Baayen, & Schreuder, 2000).

Semantic similarity seems to play a more important role than orthographic similarity. Evidence comes, for instance, from Dutch. In Dutch, the root is spelled

differently in irregular past participles than in the citation form. For instance, while *denk*, the root of *denken* 'to think', surfaces in all present tense forms of the verb, it does not in the past participle, which is *gedacht* 'thought'. Despite the orthographic dissimilarities between the past participle and its family members, the irregular past participles elicit the family size effect (Bertram, Baayen, & Schreuder, 2000; de Jong, 2002). Similar results were obtained for Hebrew (Moscoso del Prado Martín et al., 2005). To date, no research has investigated the extent to which response times can be predicted by a family size measure that, rather than categorically distinguishing between the presence versus absence of (high degree of) form overlap, takes the *degree* of form overlap into account.

Unlike the MFRM, the NDR depends on the assumption that the family size effect is driven by both form overlap and semantic similarity among the family members. The model draws on discrimination learning to simulate the visual family size effect. The NDR is a simple two-layer network with letter unigrams and bigrams as input units, and lexical meanings as output units. In order to predict response latencies, the network needs to be trained first, that is, letter unigrams and bigrams of words (e.g., #t, th, ho, ou, ug, gh, ht, t# for 'thought') are presented together with the corresponding lexical meanings. Using the equilibrium equation of the Rescorla–Wagner model (Danks, 2003), the NDR estimates weights between the input and the output units. For predicting response latencies, the model is provided with the unigrams or bigrams of a word. The presence of a unigram or bigram activates all associated meanings, modulated by the weights obtained in the training for the connections between these input and output units. The predicted response time is determined by the activation of the given lexical meaning. According to this model, the family size effect arises because a word's morphological family members provide a consistent learning environment that helps strengthen the weights from the letter unigrams and bigrams of the word's root to the word's meaning.

3.3 The present chapter

The present chapter aims to further investigate the mechanisms underlying the family size effect. It does so by, first, investigating which family size members exactly contribute to the family size effect; second, by comparing words with different morphological structures; and finally, by contrasting visual with auditory word recognition. Doing so, the present chapter will also provide more data on the role of family members in auditory word recognition.

We will report analyses of existing data from lexical decision experiments in Dutch. We chose Dutch because of the rich body of work on the family size effect in Dutch (e.g., Moscoso del Prado Martín et al., 2005; Bertram, Baayen, & Schreuder, 2000) and because of the availability of large datasets of both visual and auditory lexical decision data. To rule out that potential differences between the auditory and the visual family size effect may be due to methodological differences, we analysed the data from the visual and the auditory lexical decision experiment using exactly the same methodology.

To find out which family members drive the family size effect, and thus to shed more light on the mechanisms underlying this effect, we defined and tested six distinct measures of family size (see Table 3.1 for an overview). First, we define *Family Size* as the "type count of words in which a given target word (or, in the case of morphologically complex words, its root) appears as a constituent" (Winther Balling & Baayen, 2008). For instance, family members that would be included in this type count in the case of the word *talking* are *backtalk*, *crosstalk*, *talking-book, talk, talked, talkative,* and *talkie*. Following the findings in the literature (e.g., Moscoso del Prado Martín et al., 2004, 2005; Perdijk et al., 2012; Mulder, Dijkstra, & Baayen, 2015), we hypothesise that this measure will yield a facilitative family size effect for visual latencies regardless of the presented word's morphological structure. In contrast, for the auditory modality, we hypothesise that the facilitative family size effect only applies for words without a prefix. In the remainder, we will refer to the measure of *Family Size* with capital letters while we refer to the number of a word's morphological family members as *family size*, in lowercase.

Second, the *Onset-Aligned Family Size* (cf. Winther Balling & Baayen, 2012) is solely based on family members that contain the presented word without the suffix in word-initial position. For instance, the onset-aligned family of *talking* consists of the words *talking-book*, *talk*, *talked*, *talkative*, and *talkie*, among others, but not *backtalk* and *crosstalk.* Note that the specification *without the suffix* in the definition implies that the onset-aligned family not only contains morphological continuations (e.g., *talking-book*) but also words starting with the same root as talking (e.g., *talk*, *talked*, *talkative*, and *talkie*). Suffixes in the presented word are ignored, because the root is assumed to be the most important cue to the morphological family.

We hypothesise that family size measures based only on the onset-aligned family members will be better predictors than family size measures based

on all family members for *auditory* latencies, where consistency between the family members with the presented word from onset onwards is assumed to be relevant (cf. Winther Balling & Baayen, 2012). In contrast, because the position of the root, which drives the family size effect, is hardly relevant for written word recognition, taking only onset-aligned family members into account should perform worse than considering the whole morphological family for visual response latencies.

Third, the *Semantic Family Size* tests the role of the family members' meanings. It is based on previous findings that (especially) family members that are semantically related to the presented word drive the family size effect (e.g., de Jong, Schreuder, & Baayen, 2003; Moscoso del Prado Martín et al., 2005). Previous researchers classified family members categorically as semantically related or not, while it has been shown that gradient measures of semantic similarity better predict all types of psycholinguistic data than categorical measures (see, e.g., Buchanan, Westbury, & Burgess, 2001; Pexman & Hargreaves, 2008; Marelli & Amenta, 2018; Marelli, Amenta, & Crepaldi, 2015; Mandera, Keuleers, & Brysbaert, 2017). We therefore developed a family size measure taking degree of semantic similarity into account. We weigh the contribution of each family member to the family size on the basis of the degree of semantic similarity between the presented word and the family member, using semantic distances derived from word embeddings (e.g., Mikolov et al., 2013). For instance, *backtalk* and *talkative* contribute more than *talkie* to the Semantic Family Size of *talking*, because, in comparison to *talkie, backtalk* and *talkative* are semantically more closely related to *talking.* Based on the literature, we hypothesise that family size measures accounting for semantic similarities will predict response latencies better than family size measures that do not. This should be the case in both visual and auditory word recognition, because the family size effect is assumed to be at least partly semantic in nature.

Fourth, the *Semantic Onset-Aligned Family Size* is a combination of the Onset-Aligned Family Size and the Semantic Family Size. Semantic Onset-Aligned Family Size is identical to the Semantic Family Size except that it is only based on the onset-aligned family members. This measure will thus test whether the family size effect is driven by onset-aligned family members, weighing their contributions based on their semantic similarities with the presented word. For instance, because of its closer semantic relationship, *talkative* contributes more than *talkie* to the Semantic Onset-Aligned Family

Size of *talking. Backtalk,* although having a closer semantic relationship with *talking* than *talkie,* does not contribute to the Semantic Onset-Aligned Family Size of *talking* at all because it is not onset-aligned with *talking.* Because we expect both onset-alignment and semantic similarities to be important for the auditory modality, Semantic Onset-Aligned Family Size should outperform all three family size measures defined above in the auditory domain. In contrast, for the visual domain, we expect that Semantic Onset-Aligned Family Size will only outperform Onset-Aligned Family Size, which does not take semantic similarities into account.

Fifth, we also tested whether the degree of form overlap between the presented word and its family members is relevant for the family size effect. Previous studies, which suggested that the visual family size effect remains unaffected by the form overlap between the presented word and its family members, relied on a categorical distinction. They differentiated between family members where the root appears the same as in the presented word and those where the root differently surfaces (cf. de Jong, Schreuder, & Baayen, 2003; Moscoso del Prado Martín et al., 2005). These studies leave open the possibility that the *degree* of form overlap plays a role. To test this possibility, we developed the *Form Overlap Family Size.* It measures the form overlap between the presented word and each family member using the Levenshtein distance. Family members with a smaller Levenshtein distance and thus a greater form overlap with the presented word contribute more to this measure. Similar to Onset-Aligned Family Size, suffixes are ignored for computing Form Overlap Family Size, because the root is assumed to be the most important cue to the morphological family. For *Form Overlap Family Size*, it is important to note that also the suffixes of the family members are irrelevant. For instance, for the presented word *talking,* the family members *talkative* and *talkie*, each contribute exactly 1.0 to the *Form Overlap Family Size*, because, without their suffixes, they are identical to the presented word without its suffix. In contrast, *backtalk* contributes less because segments have to be added to transform the root of the presented word to *backtalk*. Because the degree of form overlap is a more fine-grained measure than onset-alignment, we expect, in the *auditory* modality, where onset-alignment is expected to be relevant, family size measures considering the form overlap to outperform family size measures that do not. We hypothesise that this is also the case in the visual domain, because we assume that the family size effect is based on a combination of semantic and form similarity, which is suggested by both the MFRM and the NDR.

Finally, the *Semantic Form Overlap Family Size* combines both a weighting based on the degree of form overlap and a weighting based on the semantic similarity between the presented word and each family member. For instance, for the presented word *talking, talkative* would contribute more to the Semantic Form Overlap Family Size than both *talkie,* which has a greater semantic distance, and *backtalk*, which shows less form overlap. Semantic Form Overlap Family Size should be the best predictor of all family size predictors in the context of both the visual and the auditory lexical decision data.

In family size research, it is not always clear whether family members are lexemes or word forms (e.g., Mosco del Prado Martín et al., 2004; Mulder et al., 2014). In the present chapter, both morphologically related lexemes and morphologically related word forms counted towards a word's morphological family.

Table 3.1 *Different Family Size (FS) measures briefly described and illustrated in the context of the Dutch word* optisch *'optical/optically', which has the two morphological family members* optische *'optical/optically' and* optiek *'optics' in CELEX (Baayen, Piepenbrock, Gulikers, 1996). Because this word is simplex, it also represents the word's root. In the Example column, values in parenthesis represent each family member's contribution to the corresponding family size count.*

Variable	Description	Example
(Classical) FS	Number of morphologically related words that contain the presented word's root	*optische* (1), *optiek* (1)
Onset-Aligned FS	Number of morphologically related that contain the presented word in word-initial position; suffixes are ignored	*optische* (1), *optiek* (0)
Form Overlap FS	Number of morphologically related that contain the presented word's root; the contribution of a family member is based on its Levenshtein distance with the presented word	*optische* (1), *optiek* (0.368)
Semantic FS	Number of morphologically related that contain the presented word's root; the contribution of a family member is based on its cosine similarity with the presented word	*optische* (0.830), *optiek* (0.609)
Semantic Onset-Aligned FS	Number of onset-aligned morphologically related; the contribution of a family member is based on its cosine similarity with the presented word	*optische* (0.830), *optiek* (0)
Semantic Form Overlap FS	Number of morphologically related words that contain the presented word's root; the contribution of a family member is based on both its Levenshtein distance and its cosine similarity with the presented word	*optische* (0.830), *optiek* (0.224)

In Experiment 1, we analysed data from the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, 2010). By analysing this large-scale visual lexical decision experiment, we aimed to replicate the visual family size effect that has been reported many times (e.g., Baayen, Lieber, & Schreuder, 1997; Kuperman et al., 2009). Additionally, we aimed to compare the six family size measures that we formulated. We systematically investigated which family members drive the family size effect, while systematically distinguishing between words differing in morphological structure. In Experiment 2, we analysed data from the Biggest Auditory Lexical Decision Experiment Yet (BALDEY; Ernestus & Cutler, 2015), representing an extensive auditory lexical decision experiment that was conducted in Dutch. Our objective was to examine the impact of the family size measures we defined in the context of auditory word recognition, while again systematically distinguishing between words differing in morphological structure.

3.4 Experiment I

The data and the scripts that were used for this chapter can be downloaded from: https://doi.org/10.34973/yc4q-r262.

3.4.1 Data

Our visual data stem from the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, 2010). DLP contains response latencies from 39 native speakers of Dutch to 14,089 written Dutch content words and 14,089 written pseudowords. The words differ in word class, position of stress, and morphological structure. The pseudowords conform to the phonotactic rules of Dutch and are similar to the content words in terms of number of letters and morphological structure.

We only analysed correct responses to simplex words (henceforth *simplex words*), words with one prefix (henceforth *prefixed words*), words with one suffix (henceforth *suffixed words*), and words with both one prefix and one suffix (henceforth *double-affixed words*). The resulting constrained data set contains 308,239 responses with 137,385 responses to simplex words, 5,025 to prefixed words, 156,637 to suffixed words, and 9,192 to double-affixed words.

3.4.2 Calculation of the family size predictors

We based all family size predictors on the words incorporated in CELEX (Baayen, Piepenbrock, & Gulikers, 1996), without taking stress information into account. In order to calculate each word's Semantic Family Size, we first extracted word embeddings for the word and all its family members from a Dutch *word2vec* model (Nieuwenhuisje, 2018). Word embeddings are representations of words' meanings in the form of vectors in a multidimensional space and are based on the assumption that words with similar meanings tend to show similar co-occurrence patterns with other words in large text corpora. We calculated the cosine similarity between the presented word's vector and each family member's vector. Cosine similarity is a measure that ranges between –1 for opposite and 1 for identical vectors. We normalised the cosine similarities using a min–max scaling to the range [0,1], so that we could apply a log-transformation to this predictor; all other predictors used were also log-transformed (see below). Finally, we summed up the normalised cosine similarities between the vectors of the word and its family members. As a consequence, each family member contributed to the Semantic Family Size according to its semantic similarity to the presented word. We calculated Semantic Onset-Aligned Family Size in a similar way, but excluded the family members that are not onset-aligned.

We based the Form Overlap Family Size on the number of phone edits between a word's phonemic representation and each family member's phonemic representation. We chose the number of phone edits instead of the number of letter edits because it has been shown that phonology plays a crucial role during reading (e.g., Amenta, Marelli, & Sulpizio, 2017; Jared & O'Donnell, 2017; Perrone-Bertolotti et al., 2012). In addition, Baayen et al. (2019) found indications that readers first map letter sequences onto phonemic sequences in English, which then are mapped onto semantic representations.[1] For illustrating how we computed Form Overlap Family Size, consider the Dutch word *werken* (/ʋɛrkən/) 'working', consisting of *werk* (/ʋɛrk/) 'to work/the work', and the inflectional suffix *-en* (/ən/), and its family members *afwerkt (*/ɑfʋɛrkt/*)* 'he/she/it completes', consisting of *werk* and the inflectional suffix *-t* (/t/), and *medewerker* (/meːdəʋɛrkər/) 'employee', containing the prefix *mede- (*/meːdə/*), werk,* and the suffix *-er*. As mentioned above, suffixes are ignored when the Levenshtein distance is computed between the presented

[1] We also investigated how accurately Form Overlap Family Size predicted RTs in DLP when it was based on spelling. We refrain from further discussing this measure because it yielded the least accurate predictions in comparison to all other family size predictors.

word and a family member. Thus, the Levenshtein Distance between /**ʊɛrk**ən/ and /ɑ**fʊɛrkt**/ is two (/ən/ and /t/ are suffixes and therefore are ignored) and between **/ʊɛrk**ən/ and **/meːdəʊɛrk**ər/ it is four (the four phonemes /m/, /e/, /d/, and /ə/ are to be removed; /ən/ and /ər/ are again ignored). We emphasised the degree of form overlap by squaring the computed Levenshtein distances (e.g., $2^2 = 4$; $4^2 = 16$). To ensure that words with a smaller form overlap like *medewerker* contribute less to the Form Overlap Family Size, we inverted the emphasised Levenshtein distance (e.g., $-1 \times 4 = -4$; $-1 \times 16 = -16$). To normalise the emphasised inverted Levenshtein distance, we scaled it to the interval (0;1] by passing it to the exponential function (e.g., $\exp(-4) = 0.0183$; $\exp(-16) = 1.125 \times 10^{-7}$). The Form Overlap Family Size is the sum of these scaled numbers.

For computing Semantic Form Overlap Family Size, each family member's contribution to the Form Overlap Family Size (based on its normalised, emphasised Levenshtein distance) was multiplied with its contribution to the Semantic Family Size (based on its normalised cosine similarity). The resulting products were then aggregated. For illustration, the Dutch word *afwerkt* has a normalised cosine similarity of 0.8 with *werken* and contributes to the Form Overlap Family Size with 0.0183 (its contribution is $\exp(-4) = 0.0183$; see previous paragraph), so its contribution to the Semantic Form Overlap Family Size of *werken* is 0.0147 ($0.8 \times 0.0183 = 0.0147$).

All family size predictors were log-transformed to ensure that they were normally distributed. In addition, this had the advantage that the family size predictors and the control variables are on the same scale, which makes it easier to interpret the results.

3.4.3 Generalised Additive Mixed Models (GAMMs)

The response times (RTs) were analysed with *Generalised Additive Mixed Models* (GAMMs), for which we used R 4.0.5 (R Core Team, 2021) and the package *mgcv* (Wood, 2015). GAMMs share similarities with linear mixed-effects models (LMMs; Bates et al., 2015) in that both function as regression models, capturing fixed effects as well as random intercepts (e.g., participant-specific intercepts) and random slopes (e.g., adjustments of the general effect of trial number by participant). An advantage of GAMMs over LMMs is their ability to model non-linear relationships between the dependent variable and the predictors. GAMMs can autonomously determine the shapes of these non-linear relationships, eliminating the need for a predefined specification by the researcher.

Our choice to employ GAMMs for analysing RTs aligns with Baayen et al.'s (2017) findings, demonstrating that GAMMs provide a superior fit to RTs in BALDEY compared to LMMs. Regarding the RTs of DLP, Heitmeier, Chuang & Baayen (2023) have reported non-linear effects for trial number, frequency, and word length – predictors we consider in this chapter as well. GAMMs have been applied in various linguistic disciplines, for instance in sociolinguistics (e.g., Wieling et al., 2011, 2014), phonetics (e.g., Wieling et al., 2016; Chuang et al., 2021), psycholinguistics (e.g., Lõo et al., 2018), and computational linguistics (e.g., Heitmeier, Chuang, & Baayen, 2023). In what follows, we will briefly introduce the basic concepts of GAMMs. For a more thoroughly introduction of GAMMs, see, for instance, Baayen et al. (2017).

GAMMs model the shape of the relationship between the dependent variable and predictors by using so-called basis functions. These basis functions are ordered along their degree of non-linearity. For instance, the first basis function is a horizontal line; the second basis function is a straight line with a slope; and the third basis function is a parabola. The basic functions are combined into so-called splines, which are non-linear functions; the default spline is the so-called *thin plate regression spline*. When combining the basis functions into a spline, each basis function's contribution to the final spline is weighted by multiplying the basis function with a coefficient. For instance, for modelling a horizontal line (i.e., an intercept), the first basis function would fully contribute, whereas the second and higher order basis function would receive a weight of zero. In the course of combining basis functions into splines, GAMMs penalise unnecessary non-linear splines, more explicitly, GAMMs minimise both prediction errors (like LMMs do) and the degree of non-linearity of the fitted splines. As a consequence, GAMMs only model complex non-linear relationships between the dependent variable and the predictors when the data support such a relationship. Conversely, when the data support a linear relationship, GAMMs yield a linear relationship.

Table 3.4 depicts an example of the summary of a GAMM. Part A (*parametric coefficients*) reports the effects sizes, their standard errors, t-values, and p-values of the four levels of the categorical predictor *morphological structure* ("prefixed", "suffixed" and "double-affixed", and the intercept represents the reference level "simplex"). Part B (*smooth terms*, which refers to splines) shows information about the continuous predictors. Four statistics are reported for each predictor: The *effective degrees of freedom* (edf) are a summary statistic that reflects the degree of non-linearity of the relationship between

the dependent variable and the predictor. An edf of 1 represents a linear effect, an edf of 2 represents a quadratic effect, and higher edfs represent higher degrees of a polynomial. The edf can also be a decimal number. For instance, an edf of 1.2 indicates a slightly curved line. The reference degree of freedom (*Ref.df*) is used for computing the *F-value*, which is the test statistic used for determining the *p-value*. The p-value for smooth terms indicates whether a spline significantly differs from 0 (i.e., a horizontal line). To determine the direction and the shape of the effect, it is crucial to plot the effect and to visually inspect this plot. The final three rows in Table 3.4 show the random slopes of the continuous predictors by participant.

In GAMMs, different value intervals of a predictor may not only yield different estimated effects, but also different estimated standard errors. Typically, value intervals with many observations have small standard errors while intervals with few observations have large standard errors. Obviously, these differences in standard errors between the value intervals have to be taken into account in the interpretation of the overall effect of a predictor. They are typically displayed in plots showing the relationship between the predictor and the dependent variable. We only interpret GAMMs' effects plots for intervals with a sufficient density of observations.

3.4.4 General description of the analyses

All models predict the *inverse response time* (RTinv) to the target word multiplied with −1000 (i.e., −1000/RT; Baayen & Millin, 2010). To make the predictions easier to interpret, RTinv was centred and normalised using a z-transformation.

To test the significance of our predictors, we first built baseline models that included the predictor of interest *morphological structure* (MorphStr) with the levels "prefixed", "simplex", "suffixed", and "double-affixed". In addition, the baseline model contained the most important variables that are known to predict lexical decision response latencies from visual lexical decision experiments in order to decrease the variance in the data. We then built experimental models that also contained a family size measure and investigated whether these experimental models better fitted the data and to what extent.

We compared whether the effect of family size varies between different morphological structures (MorphStr) by using treatment coding and systematically alternating the reference level of MorphStr (Wieling, 2018).

That is, first, we conducted pairwise comparisons between on the one side the level "simplex" and on the other side one of the levels "prefixed", "suffixed", and "double-affixed". Next, we set "prefixed" as reference level and compared it to either "suffixed" and "double-affixed". Finally, we set "suffixed" as reference level and compared it to "double-affixed". Based on the plots produced by the *plot_diff()* function from the package *itsadug* (van Rij et al., 2022), we determined the family size value intervals, for which differences are significant.

3.4.5 Control variables in the baseline model

Our first control variable is the *Moving Average Response Time* (maRT; ten Bosch, Ernestus, & Boves, 2018), which measures how quickly participants on average responded in the ten previous trials. This is an important control variable as the average RT *locally* fluctuates in the course of an experiment due to changes in attention (e.g., Baayen, Wurm, & Aycock, 2007; Baayen et al., 2017). Second, to capture whether participants *globally* adjusted to the task over the course of the entire experiment we included the number of each *Trial* in each session as a control variable (e.g., Ernestus & Cutler, 2015).

Third, as shorter words can be rejected or accepted earlier than longer words, we factored in word *Length* in number of letters (e.g., New et al., 2006). Fourth, the earlier a word can be differentiated from all other words in a language, the more quickly it can be recognised, which we control for with the *Form Identification Point in number of Letters* (FIPletter), which is based on the *Orthographic Uniqueness Point* (e.g., Miller, Juhasz, & Rayner, 2006). Finally, the *Frequency* of occurrence of a word is also taken into account by means of each word's form frequency as provided in CELEX (Freq; Baayen, Piepenbrock, & Gulikers, 1996).

Table 3.2 *Correlation between the control predictors in DLP.*

	maRT	Trial	Freq	Length	FIPletter
maRT	–	.00	.00	.03	.04
Trial		–	.00	.00	.00
Freq			–	−.24	−.19
Length				–	.92
FIPletter					–

All control predictors (and family size predictors, see above) were first log-transformed and then scaled and centred using a z-transformation. As can be seen in Table 3.2, Length and FIPletter are correlated with r = .92. To avoid collinearity, the two predictors were decorrelated with a principal component analysis (e.g., Farrar & Glauber, 1967; Tomaschek, Hendrix, & Baayen, 2018), resulting in the new predictors PC1 and PC2. PC1 explains 96% of the variance and equally strongly correlates with Length (r = .98) and FIPletter (r = .98). PC2 explains 4% of the variance and weakly correlates with Length (r = .2) and FIPletter (r = −.2). PC2 was not considered for further analysis, because including the least-important PC may reintroduce collinearity problems (Belsley, Kuh, & Welsch, 2005). The correlations between the resulting control predictors and the family size predictors can be seen in Table 3.3.

Table 3.3 *Correlations between the control predictors and the family size predictors in DLP.*

	maRT	Trial	Freq	PC1
Family Size	.01	.01	.36	.03
Onset-Aligned Family Size	.00	.00	.18	.00
Semantic Family Size	.01	.01	.37	.04
Semantic Onset-Aligned Family Size	.00	.00	.17	.02
Form Overlap Family Size	.00	.00	.32	.01
Semantic Form Overlap Family Size	.00	.00	.32	.01

3.4.6 Fitting of the baseline model

For fitting the baseline model, we followed the guideline provided by Bates et al. (2015) with the necessary adjustments for fitting GAMMs instead of linear mixed-effects models: We included thin plate regression splines for all control predictors and a parametric term for MorphStr. All predictors were accompanied by by-participant intercepts and by-participant random slopes. We refrained from including by-stimulus intercepts and slopes, which often results in invalid GAMMs, because of a too high concurvity with predictors based on a stimulus' characteristics like Length (cf. Baayen & Linke, 2020). We simplified the model by removing non-significant effects. Thereafter, we established whether sets of predictors had a concurvity greater than 0.7. We chose 0.7 as threshold because with this threshold half of the variance explained by a given predictor is in fact explained by other predictors. If a set of predictors had a concurvity greater than 0.7, we investigated whether the removal of one of the two predictors affected the other predictor's significance level (in terms of the p-value) or shape (based on visual inspection of an

effects plot), which indicates that the model cannot reliably estimate the two predictors' effects (Tomaschek, Hendrix, & Baayen, 2018). For instance, instead of estimating a facilitative effect, the model may estimate an inhibitory effect. If removal of a predictor affected the other predictor's significance level or shape, we removed the predictor with the larger p-value. Because the two datasets that we analyse (DLP and BALDEY) differ in many respects (e.g., modality, length of the session, number of observations), they may support different control variables, and therefore the two corresponding baseline models may be different.

3.4.7 Fitting of the experimental models

In order to investigate how well each of the six family size predictors contributed to explaining the variance in the RTs, we enriched the baseline model with each of the family size predictors, resulting in six models. We included thin plate regression splines for the family size predictor. For each family size measure, we also tested whether it interacts with MorphStr, resulting in six additional models. Thus, in total twelve experimental models were fitted.

3.4.8 Comparison of models

We performed multiple comparisons between models. First, we compared each experimental model with a simple effect of the family size measure with the baseline model. Second, we compared each experimental model containing an interaction between a family size measure and MorphStr with the experimental model just containing the simple effect of the same family size measure. We compared these nested models with likelihood ratio tests as implemented in the *itsadug* package (van Rij et al., 2017).

Finally, we compared models that only differed with respect to the family size predictor included. These comparisons were based on the *Akaike Information Criterion* (AIC; Akaike, 1974). The AIC takes into account how accurately a model fits to the data, penalising unnecessary complex models. The lower the AIC, the better the model generalises to unseen data. We computed normalised AICs by subtracting for each model its AIC from the baseline model's AIC. Although a difference in AIC of two can be interpreted as significant (Burnham & Anderson, 2004), we also computed evidence ratios that indicate how much more likely one model or a set of models is above another model or another set of models (Wagenmakers & Farrell, 2004).

3.4.9 Results

The maximal baseline model that converged included a parametric effect for MorphStr, splines for all control variables, and by-participant random slopes for Trial, Frequency, and PC1. Table 3.4 provides a summary of the baseline model.

The parametric effect for MorphStr is significant for all models tested. According to the baseline model, simplex words are responded to more quickly than prefixed (t = 2.685, p = .007), suffixed (t = 11.753, p < .001), and double-affixed words (t = 12.606, p < .001). Double-affixed words are responded to slower than prefixed (t = 5.793, p < .001) and suffixed words (t = 8.535, p < .001). There is no difference between prefixed and suffixed words (t = -0.693, p = .523). See Figure 3.1 for a visualisation.

The control predictors show the expected effects and can be seen in Figure 3.2. We only interpret effects for areas with relatively dense distributions of observations. Latencies are longer, the higher the maRT is. The effect of Trial suggests that participants fatigue in the course of a session, while there is also some indication for facilitation, probably due to adaptation, early in the session. Words with higher frequencies are responded to faster than words with lower frequencies. PC1 yields an inhibitory effect for words that are relatively short or have an average word length and in which the FIP occurs relatively early: these words are responded to more slowly, the longer they are and the later their FIP occurs.

Table 3.4 *Summary of the baseline model fitted to RTinv in DLP.*

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	−0.033171	0.002468	−13.440	< .001
MorphStrprefixed	0.033400	0.012440	2.685	0.007
MorphStrsuffixed	0.041317	0.003515	11.753	< .001
MorphStrdouble-affixed	0.120380	0.009549	12.606	< .001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(PC1)	3.972	4.000	47.58	< .001
s(Freq)	8.168	8.784	304.12	< .001
s(maRT)	11.010	13.440	6667.40	< .001
s(Trial)	8.314	8.870	40.86	< .001
s(PC1, participant)	34.907	38.000	21.58	< .001
s(Freq, participant)	35.983	38.000	28.45	< .001
s(Trial, participant)	35.164	38.000	12.36	< .001

Table 3.5 shows the results for the comparisons between the baseline model and each of the experimental models with a family size predictor. Adding a family size predictor to the baseline model significantly improves the model fit, suggesting that each family size predictor significantly predicts response latencies to stimuli in DLP. These enriched models improve further when family size is enabled to interact with MorphStr, as indicated by the difference in AICs between corresponding models with and without the interaction (see Table 3.6, Figure 3.3, and below).

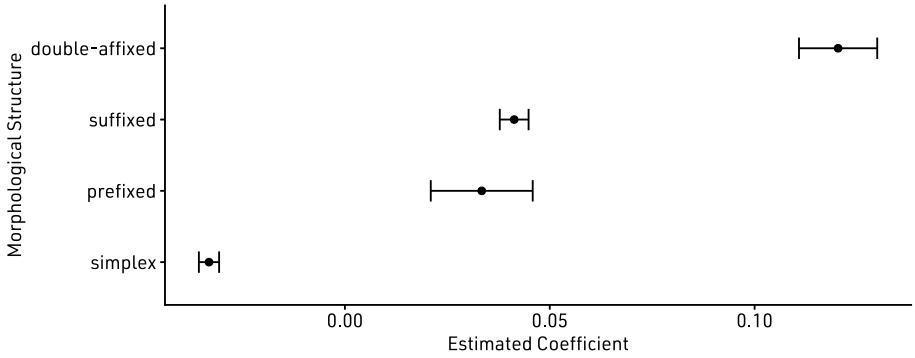


Figure 3.1 Estimated coefficients in DLP for the different levels of MorphStr in the baseline model.

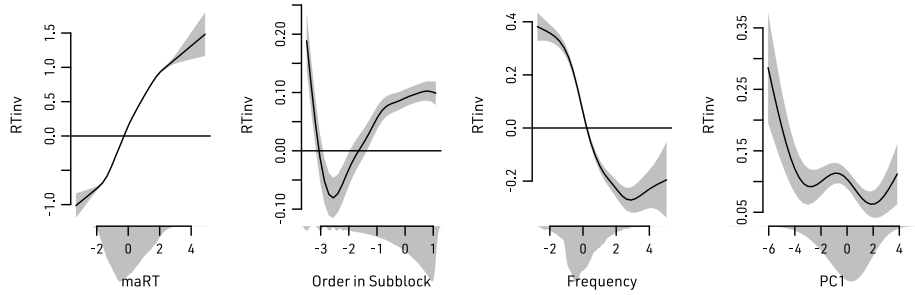


Figure 3.2 The effect of the control predictors (log-transformed, scaled and centred; x-axis) on RTinv (centred; y-axis) in the baseline model. The shaded areas indicate the 95%-confidence intervals. The density plots below the x-axis indicate the number of responses to words with the corresponding predictor value.

Table 3.5 *Comparison of the experimental models with simple effects of the family size predictors with the baseline model for DLP. The normalised AIC is the AIC of the model minus the baseline model's AIC.*

Predictor	$\chi^2(5)$	p	Normalised AIC
Family Size	1926.805	< .001	−3867
Onset-Aligned Family Size	1470.560	< .001	−2964
Semantic Family Size	2004.874	< .001	−4029
Semantic Onset-Aligned Family Size	1551.613	< .001	−3125
Form Overlap Family Size	2058.487	< .001	−4131
Semantic Form Overlap Family Size	2145.353	< .001	−4304

Table 3.6 *Comparison of the experimental models in which the family size predictor does and does not interact with morphological structure for DLP. The normalised AIC is the difference in AIC between the two models sharing their family size predictor.*

Predictor	$\chi^2(6)$	P	Normalised AIC
Family Size	221.981	< .001	−4330
Onset-Aligned Family Size	166.176	< .001	−3308
Semantic Family Size	213.907	< .001	−4481
Semantic Onset-Aligned Family Size	161.590	< .001	−3457
Form Overlap Family Size	177.747	< .001	−4512
Semantic Form Overlap Family Size	172.109	< .001	−4680

The models with family size predictors that are only based on onset-aligned family members (Onset-Aligned Family Size, Semantic Onset-Aligned Family Size) are highly unlikely compared to the models including family size predictors based on all family members (Family Size, Semantic Family Size; the two latter models are 2.175×10^{222} times more likely). Models with family size predictors that take into account the semantic similarity between each family member and the presented word (Semantic Family Size, Semantic Onset-Aligned Family Size, Semantic Form Overlap Family Size) are much more likely (2.687×10^{36} times more likely) than models with predictors that do not take the semantic similarity into account (Family Size, Onset-Aligned Family Size, Form Overlap Family Size). Finally, the models with family size predictors that weigh family members based on their degree of form overlap with the input (Form Overlap Family Size, Semantic Form Overlap Family Size) are much more likely (1.284×10^{43} times more likely) than models with predictors that do not take into account the degree of form overlap (Family Size, Semantic Family Size).

The results suggest that the visual family size effect is mostly driven by the family members that are both semantically and formally more similar to the

presented word. Accordingly, the preferred model for our data is the model including Semantic Form Overlap Family Size. A summary of this model is provided in Table 3.7.

The effect of Semantic Form Overlap Family Size is significant for simplex, prefixed, suffixed, and double-affixed words. As can be seen in Figure 3.3 (sixth row), the effect tends to be facilitative, that is, words with a larger Semantic Form Overlap Family Size are responded to faster, but this facilitative effect seems to level off for relatively big family size counts. Figure 3.3 suggests that the same pattern also holds for the other family size predictors.

The comparison between the effect of Semantic Form Overlap between different morphological structures (MorphStr) revealed that the Semantic Form Overlap Family Size effect is more pronounced for simplex words than for prefixed words ($t = 2.717$ $p = .021$), suffixed words ($t = 37.394$, $p < .001$), and double-affixed words ($t = 35.834$ $p < .001$), and that the effect for prefixed words is more pronounced than the effect for suffixed words ($t = 7.633$ $p < .001$) and double-affixed words ($t = 11.8$, $p < .001$).

Table 3.7 *Summary of the model with Semantic Form Overlap Family Size in interaction with MorphStr fitted to RTinv in DLP.*

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	−0.100057	0.002689	−37.205	< .001
MorphStrprefixed	0.045686	0.020950	2.181	0.029
MorphStrsuffixed	0.144250	0.004057	33.553	< .001
MorphStrdouble-affixed	0.146732	0.026998	5.435	< .001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(SemFormOverFS):MorphStrsimplex	8.149	8.795	450.07	< .001
s(SemFormOverFS):MorphStrprefixed	2.780	3.427	46.20	< .001
s(SemFormOverFS):MorphStrsuffixed	7.628	8.525	99.23	< .001
s(SemFormOverFS):MorphStrdouble-affixed	2.067	2.542	20.10	< .001
s(PC1)	3.912	3.995	17.65	< .001
s(Freq)	7.819	8.594	170.25	< .001
s(maRT)	11.051	13.484	6784.08	< .001
s(Trial)	8.343	8.880	42.08	< .001
s(PC1, participant)	35.018	38.000	22.71	< .001
s(Freq, participant)	36.024	38.000	29.49	< .001
s(Trial, participant)	35.221	38.000	12.63	< .001

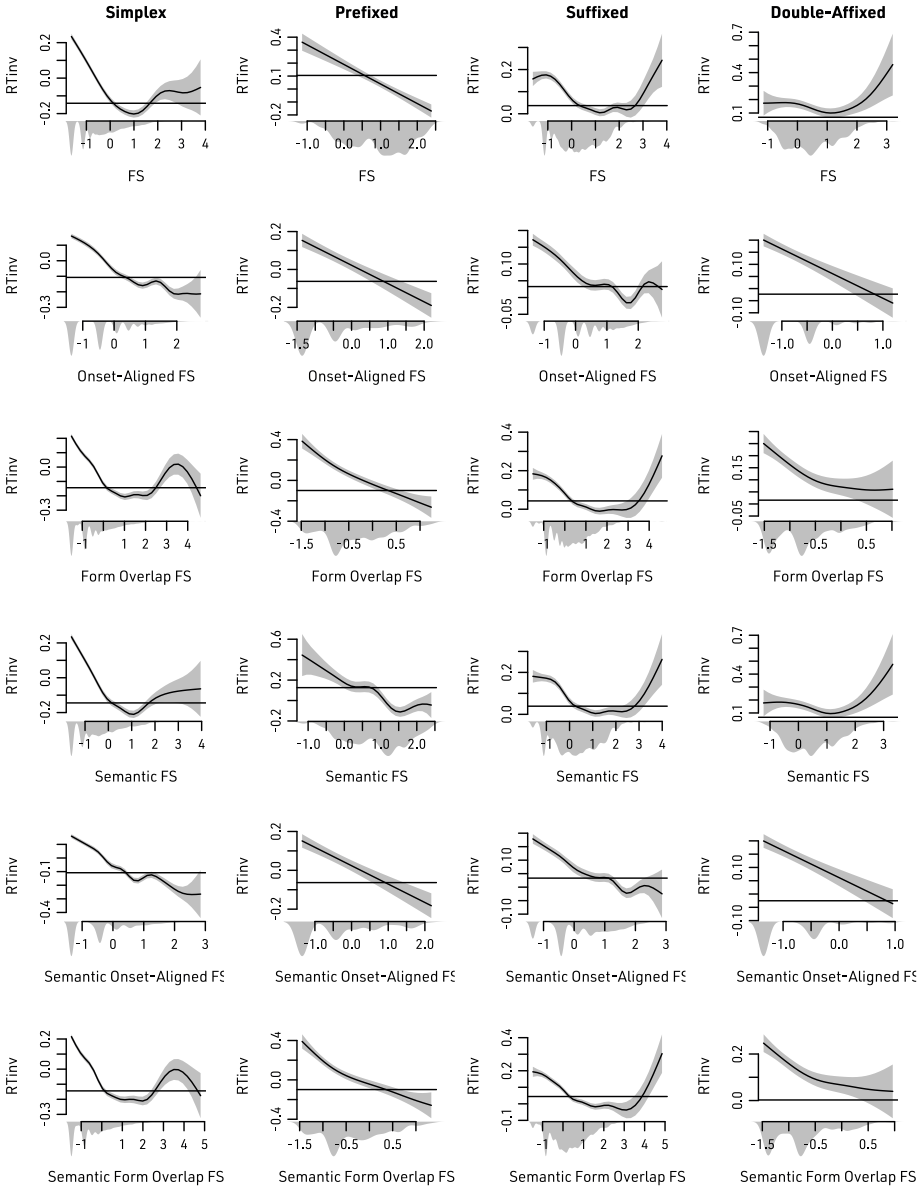


Figure 3.3 The effect of all family size predictors (log-transformed, scaled, and centred; x-axis) on RTinv (centred; y-axis) for words with different morphological structures (panels) in DLP. Horizontal lines represent the intercept for words with the corresponding morphological structure. Shaded areas indicate the 95%-confidence intervals. Density plots below the x-axis indicate the number of responses to words with the corresponding morphological structure and family size.

3.4.10 Discussion

Our results indicate that responses to Dutch written words in a lexical decision experiment can be predicted by a word's family size, regardless of how family size is operationalised, as each statistical model including a family size predictor is preferred over the baseline model. The family size effect is facilitative in general, and it seems to level off for relatively big family size counts. Family members seem to contribute more to the family size effect the larger their semantic similarity and degree of form overlap with the presented word is. On top of that, we find indications that the family size effect's size varies for words with different morphological structures.

Against the background of these results, we conclude that our methodology is appropriate to detect the family size effect. In experiment 2, we apply the same method to data from an auditory lexical decision experiment.

3.5 Experiment II

3.5.1 Data

Our auditory data stem from the *Biggest Auditory Lexical Decision Experiment Yet* (BALDEY; Ernestus & Cutler, 2015). BALDEY contains response latencies from 20 native speakers of Dutch to 2,780 spoken Dutch content words and 2,761 pseudowords. The words differ in word class, position of stress, number of syllables, and morphological structure. The content word set comprises 511 simplex words, 609 words consisting of one stem and one inflectional suffix, 770 words consisting of one stem and one derivational affix, 370 words consisting of one stem and two affixes (derivational and inflectional), 375 two-stem compounds, and 145 two-stem compounds with an inflectional suffix. Each pseudoword was created by substituting one or two segments of a real word in the experiment to make sure that the morphological and phonological structure was balanced across the word and pseudoword set.

We only analysed correct responses to simplex, prefixed, suffixed, and double-affixed words. The resulting constrained data set comprised 20,493 responses with 8,504 responses to simplex words, 739 responses to prefixed words, 14,074 responses to suffixed words, and 1,127 responses to double-affixed words. We excluded 38 (0.002%) responses from Participant 1 in Session 8 due to an encoding error and 395 responses (1.62%) that were given before the offset of the word.

3.5.2 Analysis of the data

The response latencies were analysed as described in Experiment 1 unless stated otherwise.

3.5.3 Control variables in the baseline model

The same predictors were used as in Experiment 1, except for Length and FIPletter, which were replaced by *Duration* in ms (e.g., Ernestus & Cutler, 2015) and the *Form Identification Point in ms* (FIPms; Ernestus & Cutler, 2015; Tucker et al., 2019), which is based on the *Auditory Uniqueness Point* (e.g., Marslen-Wilson, 1980). The predictors were again log- and z-transformed. Correlations between the control predictors can be seen in Table 3.8.

Due to the strong correlation between Duration and FIPms (r = .79), these predictors were decorrelated with a principal component analysis. PC1 strongly correlates with Duration (r = -.95) and FIPms (r = -.95) and explains 89% of the variance. PC2 explains 11% of the variance and less strongly correlates with Duration (r = .33) and FIPms (r = −.33). Like in Experiment 1, PC2 was not considered for further analysis. Correlations between the control variables and the family size measures are listed in Table 3.9.

Table 3.8 *Correlations between the control predictors in BALDEY.*

	maRT	Trial	Freq	Duration	FIPms
maRT	–	.01	.00	.02	.01
Trial		–	.01	−.02	−.01
Freq			–	−.16	−.12
Duration				–	.79
FIPms					–

Table 3.9 *Correlations between control predictors and family size predictors in BALDEY.*

	maRT	Trial	Freq	PC1
Family Size	.00	−.03	.31	−.09
Onset-Aligned Family Size	−.01	.01	.23	.27
Semantic Family Size	.00	−.03	.32	−.09
Semantic Onset-Aligned Family Size	−.01	.01	.23	.25
Form Overlap Family Size	.00	-.03	.27	.00
Semantic Form Overlap Family Size	.00	-.03	.27	.00

3.5.4 Results

The maximal baseline model that converged included the parametric term for MorphStr, splines for all predictors, a by-participant random intercept and by-participant random slopes for Trial and PC1. A summary of the baseline model is provided in Table 3.10.

The parametric effect of MorphStr is significant for all models tested. According to the baseline model, simplex words are responded to more quickly than prefixed ($t = 3.348$, $p < .001$) and double-affixed words ($t = 10.579$, $p < .001$). Suffixed words are also responded to more quickly than prefixed ($t = 4.178$, $p < .001$) and double-affixed words ($t = 12.005$, $p < .001$). There is no difference between simplex and suffixed words ($t = -1.908$, $p = .056$). Prefixed words are responded to more quickly than double-affixed words ($t = 4.607$, $p < .001$). See Figure 3.4 for a visualisation.

The control predictors show the expected effects, which are depicted in Figure 3.5. Latencies are longer, the higher the moving average response time is. The effect of Trial suggests that participants experienced a boost in a session's beginning, probably due to adaptation to the task, and exhibited fatigue in the course of a session. Words with higher frequencies were responded to faster than words with lower frequencies. PC1 yields a linear facilitative effect, indicating that words were responded to more quickly, the shorter they are and the earlier their FIPs occur.

Table 3.10 *Summary of the baseline model fitted to RTinv in BALDEY.*

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	−0.006340	0.044716	−0.142	0.887
MorphStrprefixed	0.088804	0.026522	3.348	< .001
MorphStrsuffixed	−0.018571	0.009735	−1.908	0.056
MorphStrdouble-affixed	0.238034	0.022500	10.579	< .001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(PC1)	1.023	1.046	70.19	< .001
s(Freq)	6.689	7.774	18.90	< .001
s(maRT)	6.084	7.339	509.43	< .001
s(Trial)	6.048	7.223	7.25	< .001
s(participant)	18.760	19.000	52.36	< .001
s(PC1, participant)	18.096	19.000	22.29	< .001
s(Trial, participant)	15.356	19.000	10.30	< .001

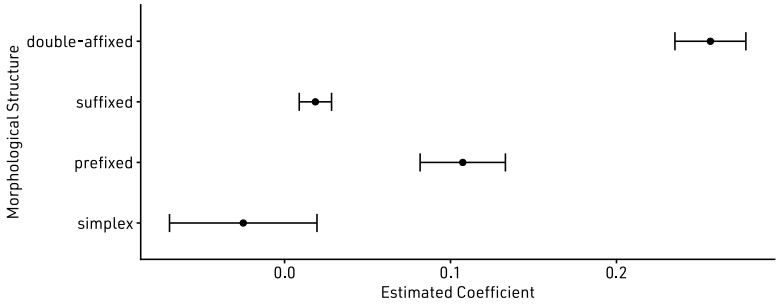


Figure 3.4 Estimated coefficients in BALDEY for the different levels of MorphStr in the baseline model.

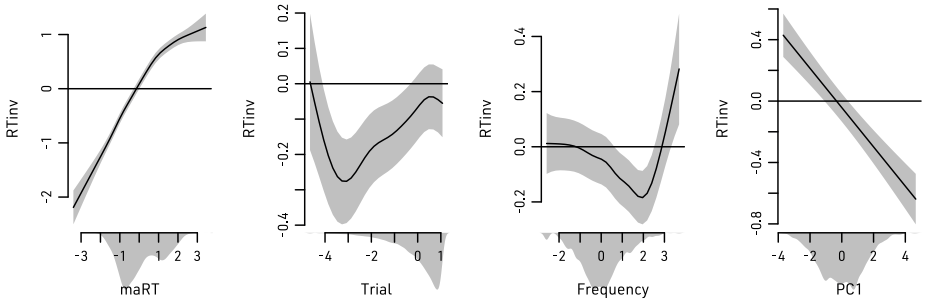


Figure 3.5 The effect of the control predictors (log-transformed, scaled and centred; x-axis) on RTinv (centred; y-axis) in the baseline model. The shaded areas indicate the 95%-confidence intervals. The density plots below the x-axis indicate the number of responses to words with the corresponding predictor value.

Table 3.11 *Comparison of the experimental models with simple effects of the family size predictors with the baseline model for BALDEY. The normalised AIC is the AIC of the model minus the baseline model's AIC.*

Predictor	$\chi^2(2)$	p	Normalised AIC
Family Size	22.279	< .001	−49
Onset-Aligned Family Size	34.676	< .001	−75
Semantic Family Size	25.589	< .001	−55
Semantic Onset-Aligned Family Size	34.356	< .001	−73
Form Overlap Family Size	39.284	< .001	−80
Semantic Form Overlap Family Size	46.639	< .001	−95

The effect of the different types of Family Size can be found in Table 3.11, which compares the AIC of each model with a family size with the AIC of the baseline model. It shows that each experimental model is significantly more

likely than the baseline model. Table 3.12 lists the effect of the interaction of each family size with the word's morphological structure on its model's AICs. It shows that each of these interactions results in a significantly better prediction of response latencies to stimuli in BALDEY, independently of the type of the family size.

Table 3.12 *Comparison of the experimental models in which the family size predictor does and does not interact with morphological structure for BALDEY. The normalised AIC is the difference in AIC between the two models sharing their family size predictor.*

Predictor	$\chi^2(6)$	p	Normalised AIC
Family Size	11.621	< .001	−58
Onset-Aligned Family Size	11.261	< .001	−89
Semantic Family Size	12.213	< .001	−65
Semantic Onset-Aligned Family Size	12.048	< .001	−90
Form Overlap Family Size	18.893	< .001	−111
Semantic Form Overlap Family Size	19.192	< .001	−127

Models including family size predictors that are only based on onset-aligned family members (Onset-Aligned Family Size, Semantic Onset-Aligned Family Size) are much more likely (556,550 times more likely) than models including family size predictors that are based on all family members (Family Size, Semantic Family Size). Models with family size predictors that weigh the semantic similarity between the presented word and each family member (Semantic Family Size, Semantic Onset-Aligned Family Size, Semantic Form Overlap Family Size) are much more likely (2,237 times more likely) than models with family size predictors that do not weigh the semantic similarity (Family Size, Onset-Aligned Family Size, Form Overlap Family Size). Finally, models with predictors that weigh the contribution of each family member based on its degree of form overlap with the presented word (Form Overlap Family Size, Semantic Form Overlap Family Size) are much more likely (3.053×10^{13} times more likely) than models with predictors that do not weigh the contribution based on the degree of form overlap (Family Size, Semantic Family Size). The model including Semantic Form Overlap Family Size fits our data the best in terms of AIC. Table 3.13 provides a summary of this model.

Table 3.13 *Summary of the model with Semantic Form Overlap Family Size in interaction with MorphStr fitted to RTinv in BALDEY.*

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	−0.04303	0.04481	−0.960	0.337
MorphStrprefixed	0.13510	0.02680	5.042	< .001
MorphStrsuffixed	0.03153	0.01070	2.947	0.003
MorphStrdouble-affixed	0.22074	0.04326	5.102	< .001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(SemFormOverFS):MorphStrsimplex	1.000	1.001	115.885	< .001
s(SemFormOverFS):MorphStrprefixed	1.000	1.001	1.034	0.309
s(SemFormOverFS):MorphStrsuffixed	2.692	3.260	7.538	< .001
s(SemFormOverFS):MorphStrdouble-affixed	1.000	1.000	1.931	0.165
s(PC1)	1.004	1.008	71.700	< .001
s(Freq)	6.481	7.591	8.992	< .001
s(maRT)	6.105	7.359	511.081	< .001
s(Trial)	6.075	7.249	7.217	< .001
s(participant)	18.761	19.000	52.779	< .001
s(PC1, participant)	18.0.99	19.000	22.382	< .001
s(Trial, participant)	15.346	19.000	10.053	< .001

The effects of different family size measures on the words with different morphological structures can be seen in Figure 3.6. The effect of Semantic Form Overlap Family Size can be seen in the sixth row. The measures have similar effects on simplex, affixed, and double-affixed words. We interpret the effect only in high density areas. The effect is significant and facilitative for simplex words, that is, simplex words with a larger (Semantic Form Overlap) Family Size are responded to faster. The same holds for suffixed words with relatively small family sizes. Due to the small number of words with large families, it is unclear whether the effect is also facilitative for these words or inhibitory. For words that contain a prefix, that is prefixed and double-affixed words, family size only tends to be facilitative if onset-alignment or form overlap is taken into account but is not significant.

Post-hoc comparisons of the effect of Semantic Form Overlap Family Size between all possible pairs of morphological structures (MorphStr) revealed that Semantic Form Overlap Families larger than 0.47 elicit a more pronounced effect for simplex words than for suffixed words ($t = 10.467$; $p < .001$). There are no significant differences in effect sizes for the comparison of other levels of MorphStr.

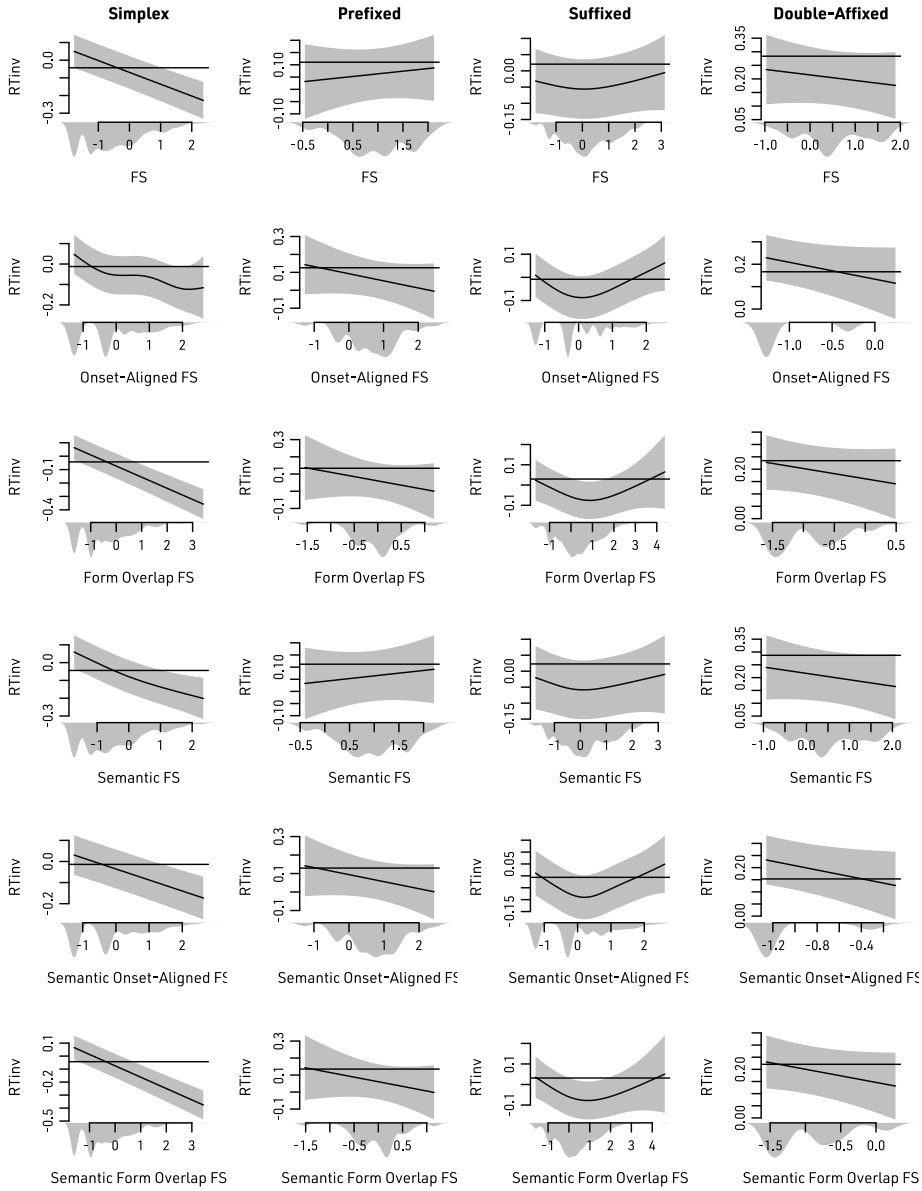


Figure 3.6 The effect of all family size predictors (log-transformed, scaled, and centred; x-axis) on RTinv (centred; y-axis) for words with different morphological structures (panels) in BALDEY. Horizontal lines represent the intercept for words with the corresponding morphological structure. Shaded areas indicate the 95%-confidence intervals. Density plots below the x-axis indicate the number of responses to words with the corresponding morphological structure and family size.

3.5.6 Discussion

Our analyses show that the family size effect is also present and facilitative in auditory word recognition for simplex and for suffixed words. For words carrying a prefix, the effect tends to be facilitative, but is not significant. Because only family size measures that take (the degree of) form overlap between family members and the presented word into account show at least in tendency the expected facilitation for prefixed and double-affixed words and a significant facilitation for simplex and suffixed words, it appears that the auditory family size is at least partly driven by form overlap. In addition, the contribution of each family member seems to depend on its semantic similarity with the presented word.

3.6 General discussion

This chapter investigates the effect of family size on both visual and auditory word recognition to shed light on the mechanisms underlying the family size effect and to examine whether there are differences between the modalities. For doing so, we analysed visual lexical decision data from DLP (Keuleers, Diependaele, & Brysbaert, 2010) and auditory lexical decision data from BALDEY (Ernestus & Cutler, 2015).

3.6.1 Similarities and differences between the visual and auditory family size effect

The family size effect is facilitative for DLP and facilitative for simplex and suffixed words in BALDEY. This was expected for DLP, because of the rich body of work on the facilitative family size effect in written word recognition (e.g., Baayen, Lieber, & Schreuder, 1997; Bertram, Baayen, & Schreuder, 2000; de Jong, 2002; Mulder, Dijkstra, & Baayen, 2015). For BALDEY, in contrast, the facilitative family size effect was less expected due to inconsistent findings with respect to the family size effect in auditory word recognition (Baayen, Wurm, & Aycock's, 2007; Winther Balling & Baayen, 2008, 2012; Wurm et al., 2006).

Family size significantly predicts latencies in DLP regardless of the presented word's morphological makeup. This is as expected. Because written words are presented at once and because of parafoveal preview (e.g., Rayner, 1998), the word's root can be immediately perceived, independently of the word's morphological structure. The root is the shared constituent across all family members and its perception is crucial for the family size effect.

In contrast, morphological structure clearly matters in BALDEY: while simplex and suffixed words yield a clear family size effect, the effect is much weaker, if present at all, for prefixed and double-affixed words. Because spoken words unfold over time, the root only becomes immediately available in simplex and suffixed words. It is precisely for these words that we found an effect of family size. The presence of a prefix might lead to a competition between words that start with the prefix. When the root is finally perceived, the competition between lexical candidates may already have progressed so far, that family members without the prefix no longer substantially influence the competition. If this is true, Onset-Aligned Family Size should facilitate responses to prefixed and double-affixed words. Figure 3.6 suggests that the effect of Onset-Alignment indeed tends to be facilitative, but our experiment probably has too few observations to show a statistically significant effect.

3.6.2 Explaining results of previous studies on the auditory family size effect

The literature on auditory family size effects reports conflicting evidence, which raises the question of how our results compare to these results. We believe that we can explain at least some of the differences between our results and those reported in the literature.

Baayen, Wurm, & Aycock (2007) investigated prefixed and suffixed words and did not find a family size effect. Similarly, Winther Balling & Baayen (2012) investigated prefixed words and compounds and found no effect. Because our data suggest that the family size effect is at best weak for prefixed words, one explanation for the null results reported is that the family size effect for suffixed words or compounds may be cancelled out by the weak effect for prefixed words.

When Winther Balling & Baayen (2012) based the family size count only on the onset-aligned family members, a facilitative family size effect emerged. At least for prefixed words, we cannot replicate this effect with our data at a level that is statistically significant. However, the effect obtained for our data goes in the right direction, suggesting that the replication fails because of too small power.

Winther Balling & Baayen (2012), in a second experiment, found an inhibitory effect for suffixed words that was driven by family members that were not onset-aligned with the presented word. It seems unlikely that the unexpected

inhibitory effect is due to deviant characteristics of Danish (compared to Dutch or English) or of the specific stimuli tested. Most likely, the facilitative effect for suffixed words did not emerge because the authors added the number of the word's morphological continuations as an additional predictor, which forms a subset of the word's family.

That Winther Balling & Baayen (2008) found a facilitative family size effect for suffixed words, but not for simplex words, is unexpected to us. Especially with respect to simplex words, our results strongly suggest the existence of a facilitative auditory family size effect.

Finally, Wurm et al. (2006) found that family size facilitated word recognition only for words with a uniqueness point that did not change when the competitors were restricted to words with the same prefix (non-CRUP words). Since our set of prefixed words contained both CRUP and non-CRUP words, our results are in line with those by Wurm and colleagues: we found that auditorily presented prefixed words may suggest a facilitative family size effect, although the effect is not statistically significant.

3.6.3 The mechanisms underlying the visual family size effect

The family size that best predicted the response latencies in the visual lexical decision experiment was the Semantic Form Overlap Family Size, which suggests that family members contribute more to the family size effect if they show more form and semantic overlap with the presented word. That semantic similarity matters, has been documented before in the literature (e.g., Schreuder & Baayen, 1997; Bertram, Baayen, & Schreuder, 2000; Moscoso del Prado Martín et al., 2004). What our results add is that this semantic similarity between each of the family members and the presented word can well be expressed as a continuous measure, based on differences in word embeddings. Recent studies indicate that humans are sensitive to gradient semantic similarities and that measures based on gradient distributional semantic models predict psycholinguistic data, including lexical decision latencies (e.g., Buchanan, Westbury, & Burgess, 2001; Pexman & Hargreaves, 2008; Mandera, Keuleers, & Brysbaert, 2017; Marelli, Amenta, & Crepaldi, 2015; Marelli & Amenta, 2018) and the N400 elicited by narrative speech (Broderick et al., 2018). We incorporated this finding in the formulation of our semantics-based family size definitions. We gauged the semantic similarity utilising a data-driven approach, which gradually measures semantic similarity, whereas previous studies on the role of semantic similarities in the context of the

family size effect differentiated between semantically related or not related (Bertram, Baayen, & Schreuder, 2000; Moscoso del Prado Martín et al., 2004; Schreuder & Baayen, 1997).

Unlike what we found, previous studies on the visual family size effect have claimed that the degree of orthographic overlap is irrelevant (e.g., de Jong, Schreuder, & Baayen, 2003; Moscoso del Prado Martín et al., 2005). They reported that Dutch irregular past participles, in which the root surfaces differently than in most of their family members, and Hebrew words, whose family members, in comparison to Dutch and similar languages, only relatively seldomly completely overlap in form, exhibit family size effects. A possible reason for this discrepancy is that we treated form overlap as a continuous rather than as a binary feature of family members. The effect of our continuous measure suggests that family members showing a small form overlap with the presented word may affect the word's recognition, but that their effects are relatively small compared to those of family members showing more overlap.

Addressing both form overlap and semantic similarity, Semantic Form Overlap Family Size bears some similarities with the measure *Orthography–Semantics Consistency* (OSC; Amenta, Crepaldi, & Marelli, 2020; Marelli, Amenta, & Crepaldi, 2015) that has been found to predict visual word recognition latencies. The OSC measure is based on words that start with the presented word, the so-called *orthographic relatives*. For instance, the orthographic relatives of the word *whisk* are *whisky, whiskey, whisker,* and *whiskered* (Marelli, Amenta, & Crepaldi, 2015). For computing OSC, the orthographic relatives' frequencies are multiplied with their cosine similarities with the presented word. The resulting values are aggregated and divided by the orthographic relatives' aggregated frequencies (normalisation). OSC is very similar to Semantic Form Overlap Family size, by taking form overlap and semantic similarity into account, but also shows substantial differences. Semantic Form Overlap Family Size computes form overlap based on phonemic representations; it is only based on morphological relatives; it is insensitive to the frequencies of the family members; and it gradually gauges form overlap (with the Levenshtein distance). These differences raise questions including why the frequencies of the relatives do not play a role in family size effects whereas they do in the OSC effects and what exactly the role is of prefixes in OSC given that they do not play a role in family size effects.

Beside the questions mentioned above, the question arises as to whether and how existing computational models that simulate the family size effect in word recognition can explain our finding that Semantic Form Overlap Family Size provides the best prediction of visual response latencies. In the Morphological Family Resonance Model (MFRM), the presented word (e.g., *preassemble*) activates the word's semantic representation (e.g., the meaning of the root *assemble*). The semantic representation then propagates activation to the family members' representations (e.g., *assembly, disassemble, disassembled, preassembled*), which share the word's root and thus its semantic representation. Then, family members' representations propagate activation via the shared semantic representation with the presented word to the presented word's representation. As a consequence, words receive the more spreading activation, the more family members they have, resulting in faster recognition. The present version of MFRM cannot simulate gradual effects of form and semantic overlap of each family members with the presented word, which would be necessary to explain our finding that Semantic Form Overlap Family Size is the best predictor in the visual domain.

In contrast, the Naïve Discriminative Reader (NDR) can account for the gradient effects of both form and meaning overlap of each family member with the presented word on its recognition. The model's input consists of unigrams or bigrams. Each unigram or bigram is connected to each lexical meaning, but the connections differ in their weight. The larger the form and semantic overlap among the family members, the stronger connections will be between this form and meaning. As a consequence, the input cues shared among semantically related family members form a stronger predictor of the meaning of the morphological family, resulting in shorter response latencies. Slightly different forms and slightly different meanings will result in slightly fewer and weaker connections, resulting in longer latencies and thus a weaker family size effect.

3.6.4 The mechanisms underlying the auditory family size effect

Given that the recognition of written and spoken words involves the same cognitive architecture, it can be assumed that the same mechanisms underly the visual and the auditory family size effect. In accordance with this, the response latencies in the auditory domain were best predicted by the Semantic Form Overlap Family Size, as was also the case for the visual modality. That is, family members are more influential in word recognition the more closely they resemble the presented word in both its form and meaning, independent of the modality.

We nevertheless also found differences between the modalities, which result from how the signal is perceived: as a whole in visual word recognition or piece by piece as time unfolds in auditory word recognition. Most importantly, contrary to our findings for visual word recognition, the family size effect was modulated by the morphological structure of the presented word. That is, we found no significant family size effects for words that contain a prefix. As mentioned above, in spoken prefixed words, the root only becomes available relatively late, compared to spoken simplex and suffixed words. The presence of a prefix might lead to a competition between words that start with the prefix.

In addition, for the auditory modality, we found that Onset-Aligned Family Size is a better predictor than Family Size. This indicates that, in auditory word recognition, family members sounding like the presented word for a certain duration are more important for the family size effect than family members showing overlap in sound only later. The fact that, nevertheless, the best predictor is Semantic Form Overlap Family Size, suggests that the binary distinction between onset-aligned and non-onset-aligned family members may be too simple for spoken word recognition. For illustration, consider the Dutch word *optisch* 'optical/optically' with the family members *optische* 'optical/optically' and *optiek* 'optics' (see also Table 3.1). While both *optische* and *optiek* show a large degree of form overlap with *optisch,* only *optische* would contribute to the Onset-Aligned Family Size. Future research should test a family size measure that not only weighs form and meaning overlap in a gradient way, but also to what extent family members share their onsets.

One of the models that can simulate the visual family size effect – MFRM – has never been applied to auditory stimuli. It takes the whole presented word as its input and provides a predicted response time as its output. For simulating the auditory family size effect, it seems crucial to incrementally take the spoken word as input and simulate how the lexical competition unfolds over time. This would better simulate the way listeners perceive words and it would be necessary to explain the difference between words containing a prefix on the one hand and simplex and suffixed words on the other hand.

With respect to the Naïve Discriminative Reader, it is noteworthy that NDR is the predecessor to the *Naïve Discriminative Learning* (NDL) and *Linear Discriminative Learning* (LDL) model (e.g., Chuang & Baayen, 2021). Shafaei-Bajestan et al. (2021) demonstrated that at least LDL can incrementally take portions of the auditory signal as input and by doing so can model auditory

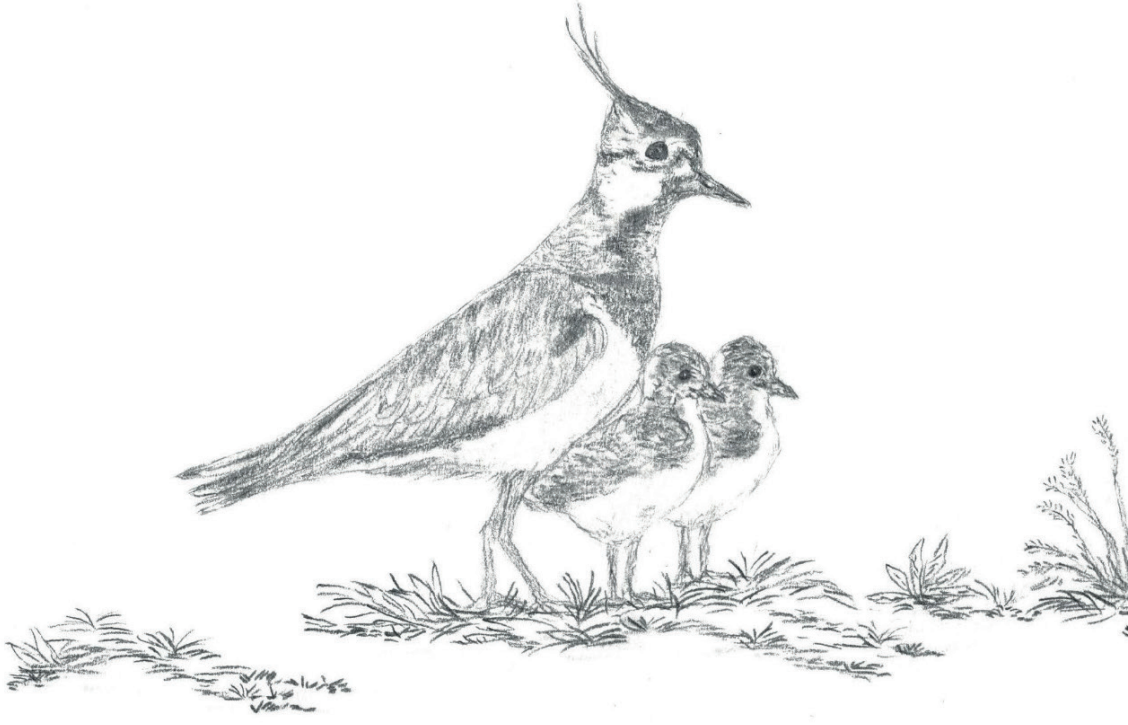
word recognition, which seems to be a necessity for modelling the auditory family size effect. Future research should aim to demonstrate that LDL can simulate our results.

3.7 Conclusion

The present chapter successfully replicated the visual family size effect, that is, it showed that words with larger morphological families are responded to faster in visual lexical decision tasks. Our results indicate that non-prefixed words with larger morphological families are also responded to more quickly in auditory word recognition. In contrast, prefixed words do not show a statistically significant family size effect in auditory word recognition. That is, while a word's morphological structure is irrelevant for the emergence of family size effects in visual word recognition, it is relevant in auditory word recognition. This is likely due to the fact that, in visual word recognition, all segments of a word are available at the same time, while this is not the case in auditory word recognition. This finding can well explain the discrepancies that exist between the few studies reported in the literature investigating family size effects in auditory word recognition.

We compared the effects of family size measures that differ in how much the different family members contribute in order to shed more light on the underlying mechanisms. These comparisons showed that family members that are more semantically and formally related are more influential. For auditory word recognition, on top of this, the onset-aligned family members seem to be especially influential, which may again be related to the time-course with which spoken words become perceptible. The relevance of both formal and semantic properties of family members should be theoretically possible to simulate with the Naive Discriminative Reader. In order to account for the role of morphological structure in auditory word recognition, this model has to take into account that the phonemes of spoken words do not all become perceptible at the same time. Only in this way can the model account for the finding that the best predictor of visual and auditory word recognition is the Semantic Form Overlap Family Size, which factors in form and semantic overlap between the presented word and its family members in a gradient manner.

3



Chapter 4

Can the Discriminative Lexicon Model Account for the Auditory Family Size Effect?

This chapter is based on:
Müller, H., ten Bosch, L., & Ernestus, M. (2024). Can the Discriminative Lexicon Model account for the family size effect in auditory word recognition? *Nota Bene, 1*(2), 176–192. https://doi.org/10.1075/nb.00010.mul

Abstract

Words with larger morphological families elicit shorter response times (RTs) in lexical decision experiments (e.g., Bertram et al., 2000). One possible account for this family size (FS) effect draws on the Discriminative Lexicon Model (DLM; Chuang & Baayen, 2021), positing that morphological family members strengthen relationships between forms and meanings. While it has been shown that the DLM successfully explains FS effects in reading (Mulder et al., 2014), we investigated whether it does so in listening too. We trained the computational model LDL-AURIS (Shafaei-Bajestan et al., 2023), which implements the DLM, on Dutch and show that a measure derived from LDL-AURIS accounts for variance in auditory lexical decision RTs in Dutch, and also partially accounts for the same variance in the RTs as the auditory FS effect. Future research should investigate whether some other measure derived from the DLM can fully explain FS effects in listening.

A word's *family size* is the count of all word types in which the given word's root occurs as a constituent. For instance, family members of the word *think* are, among others, *doublethink, thinks, rethinking,* and *unthought*. In both visual and auditory lexical decision experiments, words with larger morphological families elicit quicker responses (e.g., Moscoso del Prado Martín et al., 2004; Chapter 3). One explanatory account for this *family size effect* stems from the theory of the *Discriminative Lexicon Model* (DLM; Chuang & Baayen, 2021). A computational implementation of the DLM successfully simulated the family size effect in the reading of English nouns (Mulder et al., 2014). It is not self-evident that the DLM can explain the family size effect in listening as well because the processes underlying written and spoken word recognition systematically differ. The present chapter investigates to what extent the DLM explains the auditory family size effect.

4.1 The family size effect

The family size effect has been observed in multiple studies and for multiple languages (e.g., Bertram et al., 2000; Mulder et al., 2014) in both reading and listening. It has been suggested that the family size effect is especially driven by semantic similarities between family members (e.g., Moscoso del Prado Martín et al., 2004). Moreover, family members contribute more to the effect the greater their similarity in form with the word to be recognised (Chapter 3).

There are fundamental differences between how the family size effect manifests itself in visual and in auditory word recognition. In visual word recognition, the morphological structure of the word to be recognised does not affect the family size effect: the effect has been documented for prefixed (e.g., Moscoso del Prado Martín et al., 2004), simplex (e.g., Mulder et al., 2013), and suffixed words (e.g., Bertram et al., 2000). In contrast, in auditory word recognition, the family size effect is elicited only by simplex and suffixed words, but not by prefixed words, suggesting that the morphological structure of the word to be recognised interacts with the effect (Chapter 3).

Differences between the visual and the auditory family size effect can be explained in light of systematic differences between reading and listening. During reading, a word's characters can all be simultaneously processed due to parafoveal preview (e.g., Rayner, 1998). This simultaneity renders it irrelevant whether the word's root, which is the shared element among the

family members, occurs as first, second, or third constituent. In contrast, in auditory word recognition, words unfold over time and the word recognition process is assumed to start as soon as the audio input becomes available. Human auditory word recognition considers all words stored in memory that are (to some extent) compatible with the audio presented so far and gradually winnows out words that become more incompatible when the audio unfolds (e.g., Marslen-Wilson & Welsh, 1978). Because of the incremental unfolding of the audio signal, prefixes are first perceived and processed, then roots, then suffixes. This may explain why the family size effect is less strong for prefixed words: the recognition process for prefixed words is already well under its way before the root, linking the members of a morphological family, becomes discernible.

4.2 Discriminative Lexicon Theory

Mulder and colleagues (2014) propose that the family size effect can be understood in terms of discriminative learning in the DLM. Discriminative learning supposes that the association between parts of a word's form and the word's meaning is strengthened when both occur together and weakened when one of them is present while the other is absent. The latter may occur, for instance, when a morpheme has several meanings or when a sound sequence (e.g., /rɛd/) is part of words with different meanings (e.g., *red*, *bread*). A stronger association leads to faster recognition when the word is presented. In the DLM, the family size effect is explained by the principle that the more family members a word has, the stronger the association between the word's root and its meaning.

The DLM is usually implemented as a two-layer neural network that takes as its input words' feature representations and predicts words' meaning representations. Feature representations can take the form of letter sequences or acoustic features, while meanings can be represented by arbitrary identifiers or semantic vectors (see below). For making predictions, the network first has to be trained, that is, it has to establish the association weights between word feature representations and meaning representations, on the basis of the input features and meanings of numerous words. When association weights have been established, the DLM can determine a given input word's meaning by comparing the support of the word's features for all word meanings in the lexicon, modulated by the association strengths.

Vice versa, the DLM can produce a word's form, given a word's meaning. Previous studies have shown that measures derived from the DLM can predict behavioural data including visual lexical decision data and acoustic durations (for an overview, see Chuang & Baayen, 2021).

Mulder and colleagues (2014) implemented an early version of the DLM as a computational model for word reading, called the *Naïve Discriminative Reader*. The Naïve Discriminative Reader takes as its input character trigrams. Mulder and colleagues used the support for the meanings of the words in the lexicon to simulate lexical decision data. Family size was a significant predictor for the observed and the simulated lexical decision data, suggesting that the DLM can explain family size effects.

There is only one implementation of the DLM for spoken word recognition: LDL-AURIS (Shafaei-Bajestan et al., 2023). This implementation takes as its input words' audio recordings, of which the frequency spectra are summarised by means of *Continuous Frequency Band Summary Features* (C-FBSFs). Whereas the Naïve Discriminative Reader represents the meaning of a word with a unique letter sequence (localist representation), LDL-AURIS represents meanings with semantic vectors produced by a distributional semantics model. These vectors reflect that words with similar meanings tend to co-occur with the same set of words.

Shafaei-Bajestan and colleagues (2023) tested LDL-AURIS by determining how well it can recognise words sliced from continuous speech. Recognising cut-out words is a difficult task for human listeners, because, in everyday speech, words tend to be coarticulated and reduced, which makes them difficult to identify when presented without their contexts. Accordingly, human participants only identified 20.8% to 44.0% of the words sliced out of their contexts (Arnold et al., 2017). To determine whether LDL-AURIS correctly recognised a presented word, Shafaei-Bajestan and colleagues (2023) computed the correlation between the semantic vector computed for this word and all vectors in the lexicon. If the vector of the correct meaning was closest to the computed vector, the presented word was assumed to be correctly recognised. LDL-AURIS recognised 16% of the words, which approximated the lower bound of human performances in this difficult task. It has yet to be investigated whether LDL-AURIS can also predict the time listeners need to recognise a word, for instance, the response times (RTs) from lexical decision experiments.

4.3 The present chapter

The present chapter focused on two research questions. First, we investigated whether LDL-AURIS can account for how quickly listeners recognise spoken words. More specifically, we investigated whether LDL-AURIS predicts RTs from an auditory lexical decision experiment. Second, we investigated whether and to what extent LDL-AURIS accounts for the same variance in the auditory lexical decision RTs as family size does.

Previous research has used different definitions of family size, resulting in different family size measures. In the present chapter, we focused on three of them. The first measure is *Classical Family Size*, for which all words including the presented word's root (i.e., family members) equally weigh. Like all family size measures, Classical Family Size yields a facilitative effect in lexical decision experiments (e.g., Bertram et al., 2000). The second measure is *Semantic Family Size*, for which the weight of a family member depends on the strength of its semantic relation to the presented word. Semantic Family Size yields better predictions for RTs than Classical Family Size (e.g., Moscoso del Prado Martín et al., 2004). Following Chapter 3, the third measure is *Semantic Form Overlap Family Size,* for which the weight of a family member depends on both the strength of its semantic relation to the presented word and its form overlap with the presented word. Semantic Form Overlap Family Size is the best predictor of all family size measures for visual and auditory lexical decision RTs (Chapter 3). For a detailed description of how we computed these family size measures, see Chapter 3.4.2.

We tested LDL-AURIS against the RTs from the *Biggest Auditory Lexical Decision Experiment Yet* (BALDEY; Ernestus & Cutler, 2015), a Dutch large-scale auditory lexical decision experiment. We chose Dutch because Chapter 3 showed that the three above-mentioned family size measures are statistically significant predictors of lexical decision RTs in Dutch.

In order to investigate our research questions, we compared three types of models, as summarised in Table 4.1. First, we built a statistical baseline model to predict the RTs from BALDEY that includes the most important control variables known to predict auditory lexical decision RTs (see Chapter 4.4.4), in order to decrease the variance in the RTs. We compared this baseline model with a model that also contained a predictor derived from LDL-AURIS (in interaction with the word's morphological structure). If the latter model is better, LDL-AURIS contributes to explaining the RTs.

Second, we produced three new statistical models by extending the baseline model with both the LDL-AURIS measure and a family size measure (both in interaction with morphological structure). We investigated whether any of these three new models better fit the RTs than an extension of the baseline model with just the LDL-AURIS measure. If so, the LDL-AURIS measure does not fully explain family size effects.

Third, we investigated whether the LDL-AURIS measure accounts for at least part of the family size effect. To this end, we investigated how much any of the three family size measures (in interaction with morphological structure) improves the model fit when added to the baseline model and compared this to how much any of the three family size measures improves the model fit when added to a model also containing the LDL-AURIS measure (in interaction with morphological structure). If the presence of the LDL-AURIS predictor results in a smaller improvement in terms of model fit, the LDL-AURIS measure accounts for at least part of the family size effect.

Table 4.1 *Overview of model comparisons and the conclusions that can be drawn based on the results.* SemDens *refers to Semantic Density, the LDL-AURIS measure that we tested.* Family Size *represents any of the three tested family size measures.*

Model 1	Model 2	Interpretation of potential results
Baseline	Baseline + SemDens	If Model 2 is better than Model 1, the LDL AURIS measure accounts for RTs.
Baseline + SemDens	Baseline + SemDens + Family Size	If Model 2 is better than Model 1, the LDL-AURIS measure does not (fully) account for family size effects.
Baseline (1a) vs. Baseline + Family Size (1b)	Baseline + SemDens (2a) vs. Baseline + SemDens + Family Size (2b)	If the difference in the model's goodness of fit with the RTs between Models 1a and 1b is larger than between Models 2a and 2b, the LDL-AURIS measure accounts at least to some extent for the family size effects.

Because LDL-AURIS has not yet been used to predict lexical decision RTs, we based the choice of our predictors on previous studies using the DLM to predict visual lexical decision RTs. Previous studies identified two predictors. The first predictor is *Target Correlation*, which is defined as the correlation between the semantic vector produced by LDL-AURIS on the basis of the audio input and the semantic vector of the correct word in the lexicon (Heitmeier et al., 2023a). Because this predictor predicted RTs with statistical significance for our BALDEY dataset only for certain random-effect structures, we refrain from further discussing this predictor.

The second predictor is *Semantic Density*, which is the average cosine similarity between the semantic vector produced by LDL-AURIS based on the audio signal and each of the ten closest semantic vectors in the lexicon, in terms of cosine similarity. Heitmeier and colleagues (2023b) report that higher Semantic Densities correlate with shorter RTs. Their explanation for this finding is that when the semantic vector produced by the model lands in areas of more words, the presented word has a higher *wordlikeness*, which facilitates a "yes" response in lexical decision experiments.

4.4 Experiment

The data and the scripts that were used for this chapter can be downloaded from: https://doi.org/10.34973/x6v3-yj45.

4.4.1 Data

We predicted the RTs from BALDEY, which contains response latencies from 20 native speakers of Dutch to 2,780 spoken Dutch content words and 2,761 pseudowords. We only analysed correct responses to all real words, except for compounds, that were also part of the training set of LDL-AURIS (see Chapter 4.4.3). The dataset thus comprised 15,936 responses with 5,908 responses to 322 unique simplex words, 227 responses to 12 unique prefixed words, 8,875 responses to 478 unique suffixed words, and 926 responses to 50 unique words containing both a prefix and a suffix. We excluded 24 (0.15%) responses from Participant 1 in Session 8 due to an encoding error and the 248 responses (1.56%) given before stimulus offset.

4.4.2 Training LDL-AURIS

We trained LDL-AURIS on the audio recordings of *Component o* of the *Spoken Dutch Corpus* (Oostdijk, 2000), which contains read-aloud speech from Dutch native speakers. We chose read-aloud speech because it is usually clearly pronounced, like the stimuli in BALDEY. Word tokens were sliced out from their acoustic context based on word segmentations as provided in the corpus. We removed mispronounced, incomplete, and unintelligible word tokens. The resulting dataset contains 550,688 word tokens (39,278 word types).

LDL-AURIS' input matrix specifies for each word token its acoustic properties in the form of C-FBSFs. The output matrix specifies for each word token its gold standard semantic vector, which is the semantic vector derived from a

distributional semantics model. We used a Dutch distributional semantics model (Nieuwenhuijse, 2018) that was trained on more than 600 million messages on Dutch social media, news, blogs, and forums, with word2vec (Mikolov et al., 2013). We removed 36,002 word tokens (14,646 word types) from the training data of LDL-AURIS, because the distributional semantics model did not provide semantic vectors for these words.

We trained LDL-AURIS in *julia* (Bezanson et al., 2017) with the package *JudiLing* (Luo et al., 2020). All parameters were exactly set as by Shafaei-Bajestan and colleagues (2023), who provide more details about the training procedure of LDL-AURIS.

4.4.3 Calculation of the family size measures

We determined all family size measures exactly as in Chapter 3. We based the family size measures on words incorporated in CELEX (Baayen et al., 1996).[2] Classical Family Size has a mean of 7.07 (SD = 0.36), Semantic Family Size has a mean of 7.19 (SD = 0.28), and Semantic Form Overlap Family Size's mean is 5.28 (SD = 0.98). For the statistical analyses, which were conducted with these measures, the measures were first log-transformed and then normalised with a z-transformation. After this pre-processing, Classical Family Size strongly correlates with Semantic Family Size (r = .998) and with Semantic Form Overlap Family Size (r = .804); Semantic Family Size and Semantic Form Overlap Family Size also strongly correlate (r = .810).

4.4.4 Control variables in the baseline model

Our baseline model included four control variables. The first is *Morphological Structure* (MorphStr) with the levels "prefixed", "simplex", "suffixed", and "double-affixed", because the auditory family size effect has been shown to vary with the morphological structure of the presented word (Chapter 3). Second, the *Moving Average Response Time* (maRT) models the weighted average response time over preceding trials (ten Bosch et al., 2018). Third, to capture a participant's adjustment to the task throughout the entire

[2] Because LDL-AURIS was trained on fewer word types than are available in CELEX, we tested whether the results of this study change when the family size measures are only based on those that also occur in the LDL-AURIS training data. These alternative family size measures correlate with the family size measures reported in this paper with coefficients between .90 and .93 for any family size measure. More importantly, these alternative measures yield results very similar to those reported in this study. We chose to present the results from the family size measures based on the complete CELEX database in this study because we believe that they better reflect a listener's knowledge of words, which is not only based on listening but also on reading.

experiment, we incorporated the number of each *Trial* (e.g., Ernestus & Cutler, 2015). Fourth, we factored in form *Frequency*, which we obtained from CELEX (Freq; Baayen et al., 1996). All control variables were first log-transformed and then z-transformed for scaling and centring.

The correlation coefficients between each pair of control variable and between each control variable and Semantic Density or a family size measure is smaller than .1, except that Frequency weakly correlates with the three family size measures (r_{min} = .34, r_{max} = .38). These correlations are considered too low to be problematic.

4.4.5 Estimation and comparison of the models

We implemented all statistical models as *Generalised Additive Mixed Models*, with R 4.0.5 (R Core Team, 2021) and the package mgcv (Wood, 2015). We preferred this type of model over *Linear Mixed-Effects Models* (e.g., Bates et al., 2015), because the former can easily detect both linear and non-linear effects, whereas the latter can only detect linear effects. A practical introduction to Generalised Additive Mixed Models is provided by, for instance, Chuang and colleagues (2021).

We fitted all models in the style of Bates and colleagues (2015). That is, for the baseline model, the initial model comprised as predictors a) a parametric term for the categorical variable Morphological Structure, b) thin plate regression splines for all continuous control variables, c) a by-participant intercept, and d) by-participant random slopes for all continuous variables. We subsequently simplified this model by step-wise elimination of predictors that did not reach statistical significance. Then, we assessed whether pairs of predictors exhibited a concurvity exceeding 0.7, suggesting that half of the explained variance attributed to a given predictor is actually accounted for by other predictors. For pairs that surpassed the threshold of 0.7, we tested whether eliminating one of the two predictors impacted the other predictor's significance level (measured by the p-value) or shape (as depicted in an effects plot). If so, we eliminated the predictor with the smaller p-value. This procedure ensures that the model can accurately estimate all included predictors' effects (Tomaschek et al., 2018).

As summarised in Table 4.1, we compared the baseline model and models extended with Semantic Density or a family size measure. Following standard procedures (e.g., Chuang et al., 2021), we did so using a χ^2 test on likelihood

scores for nested model comparisons. For comparing the difference in model improvement between two pairs of models, we compared decrease in AIC.

4.4.6 Results

Because the baseline model is not of interest by itself, we will not discuss it here. We just note that the control variables show approximately the same types of effects in the baseline model as in the best model developed in this chapter, which is summarised in Appendix E.

Adding Semantic Density in interaction with Morphological Structure as predictor to the baseline model results in a significantly better fit to the data ($\chi^2_{(2)}$ = 366.446, p < .001). Semantic Density is a significant predictor of the RTs of simplex (F = 66.237, p < .001), suffixed (F = 303.520, p < .001), and prefixed words (F = 4.891, p = .016), but not of double-affixed words (F = 0.591, p = .442). This shows that a predictor derived from the DLM as implemented in LDL-AURIS can explain variance in auditory lexical decision RTs of words that are made of at most two morphemes. As illustrated in Figure 4.1, the effect of Semantic Density is inhibitory, that is, stimuli with a greater Semantic Density are responded to more slowly. The effect seems to level-off for values of Semantic Density greater than 0.6.[3]

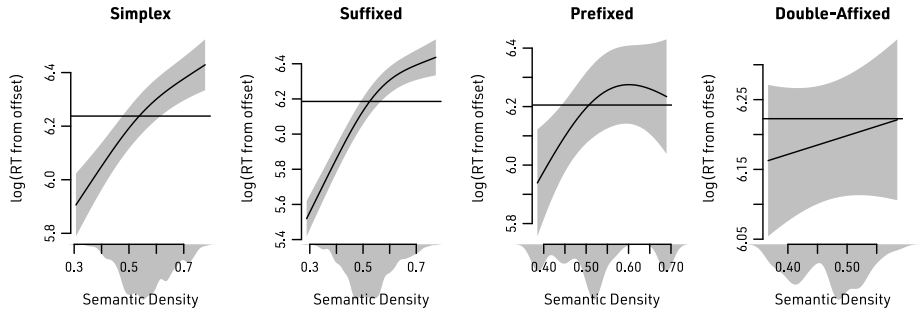


Figure 4.1 Partial effect of Semantic Density (x-axis) on log-transformed RTs (y-axis) for words with different morphological structures (panels). The density plots below the x-axes indicate the number of responses to words with the corresponding Semantic Density.

The further addition of an interaction between any of the three family size predictors by Morphological Structure to the model with Semantic Density

[3] Based on a question from a reviewer, we tested whether including an interaction between semantic density and word frequency would lead to a significantly better model, which is not the case.

results in an even better fit to the data (see Table 4.2), with Semantic Density again showing the same types of effects as illustrated in Figure 4.1. This shows that any family size measure explains variance in the RTs that is not explained by Semantic Density.

Table 4.2 *Comparisons between a) the Baseline model plus Semantic Density by Morphological Structure and b) the Baseline Model plus Semantic Density by Morphological Structure and plus any family size predictor by Morphological Structure.*

Family Size Predictor	$X^2(8.00)$	p
Classical Family Size	12.087	.002
Semantic Family Size	13.664	< .001
Semantic Form Overlap Family Size	28.680	< .001

A summary of the best model, which includes Semantic Form Overlap Family Size in interaction with Morphological Structure, is similar to the summaries of the other two models with family size predictors and can be seen in Appendix E. As shown in Figure 4.2, the effect of Semantic Form Overlap Family Size is facilitative, as expected, but the size of the effect varies with the word's morphological structure.

Finally, Figure 4.3 shows that adding a family size predictor to the baseline model improves this model more than adding a family size predictor to a model that also contains the predictor Semantic Density. These results suggest that Semantic Density partially explains the same variance in the RTs as the family size predictors.

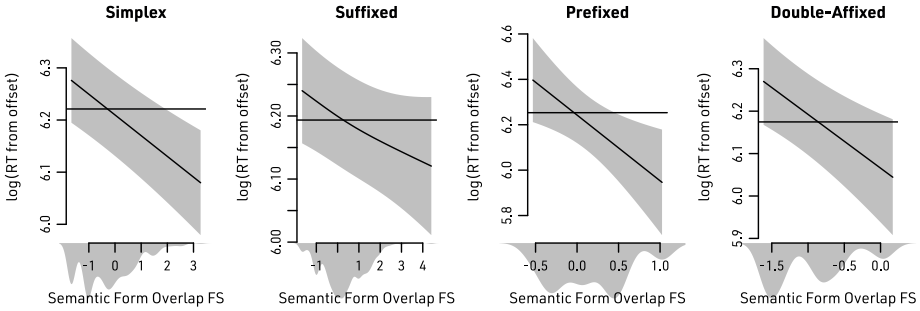


Figure 4.2 Partial effect of log-transformed and centred Semantic Form Overlap Family Size (FS; x-axis) on log-transformed RTs (y-axis) for words with different morphological structures (panels). The density plots below the x-axes indicate the number of responses to words with the corresponding Semantic Form Overlap FS.

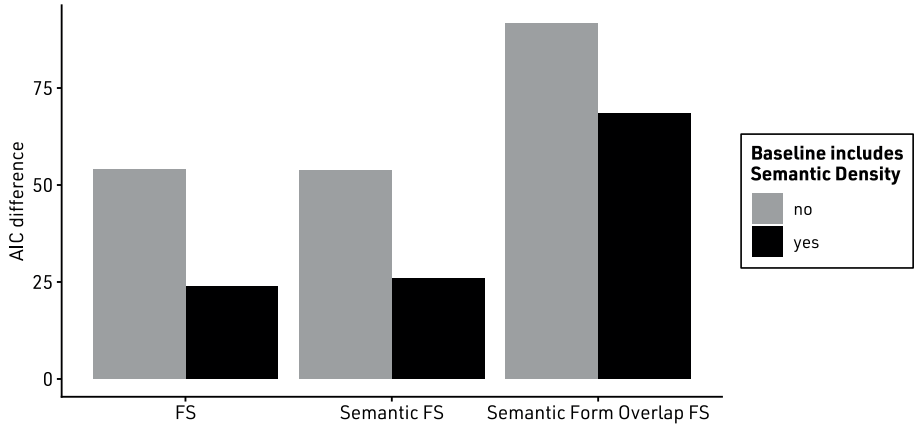


Figure 4.3 Improvement of the model's fit in terms of reduced AIC-points when either family size (FS) predictor is added to the baseline model (grey) and a model that also contains Semantic Density (black).

4.5 Discussion

This chapter has addressed the question whether the Discriminative Lexicon Model (DLM) can account for response times (RTs) from an auditory lexical decision experiment and for the family size effect in those RTs. We derived a measure, Semantic Density, from the computational model LDL-AURIS (Shafaei-Bajestan et al., 2023), which implements the DLM for auditory word recognition, and tested it against the RTs of the Dutch large-scale auditory lexical decision experiment BALDEY (Ernestus & Cutler, 2015).

First, we investigated whether Semantic Density significantly accounts for variance in the RTs, which is the case. This chapter thus enriches previous research that up to now has only shown that LDL-AURIS can recognise words approximately as accurately as human listeners (Shafaei-Bajestan et al., 2023), by showing that LDL-AURIS can also account for variance in auditory lexical decision RTs. Possibly, Semantic Density accounts for less variance in the RTs to words containing both a prefix and a suffix than to words of a simpler morphological structure due to sparseness of those complex words in LDL-AURIS' training set.

In this chapter, Semantic Density yielded an inhibitory effect. This may be surprising because Heitmeier and colleagues (2023b) reported a facilitative

effect. One explanation for the inhibitory effect is already suggested by Heitmeier and colleagues: A higher Semantic Density implies that the meaning computed by LDL-AURIS is more similar to more words in the lexicon, which may render it more difficult to identify which of the meanings in the lexicon was intended. Our finding that a greater competition between meanings results in longer RTs is in line with cohort-driven auditory word recognition models such as DIANA (ten Bosch et al. 2022). The inhibitory effect in BALDEY suggests that in this experiment, participants only accepted a word as a real word when they knew the word's meaning. In visual lexical decision experiments, participants may already have accepted a word because it was word-like, leading to a facilitative effect of Semantic Density.

Second, we investigated whether any of the three family size measures that we tested accounts for variance in the RTs that is not accounted for by Semantic Density, which is the case for all of them. Because this shows that Semantic Density does not fully account for the family size effect, we finally tested whether Semantic Density at least partially accounts for family size effects. For doing so, we tested whether adding any family size measure to a model containing Semantic Density improves the fit to the data less than adding this family size measure to a model without Semantic Density. This appeared to be the case, for all three family size measures. Therefore, our results suggest that Semantic Density at least partially accounts for the family size effects in listening. This chapter therefore expands previous research by showing that not only the visual but also the auditory family size effect can at least partially be understood in terms of discriminative learning in the DLM.

LDL-AURIS relies on associations between forms and meanings. A given association is strengthened by more word tokens supporting this association (i.e., by more family members showing form overlap). Consequently, it may be expected that LDL-AURIS is most effective in explaining the effect of Semantic Form Overlap Family Size. As mentioned above, we tested for each family size measure to what extent it improves a model with and without Semantic Density. The more the addition improves a model without Semantic Density compared to the model with Semantic Density, the more effectively Semantic Density explains the effect of this family size measure. Contrary to expectations, Semantic Density explains the effect of Semantic Form Overlap Family Size to a lesser extent than the other two family size measures' effects. Apparently, the associations between form and meaning in the DLM contain slightly different information than Semantic Form Overlap Family Size. A

probable cause is that associations between forms and meanings in the DLM are not only strengthened by word tokens supporting these associations, but also weakened by word tokens that do not support these associations, by representing similar forms but different meanings, or vice versa. Another probable cause is that LDL-AURIS is trained on word tokens whereas family size is based on word types. The association strength between a form and a meaning in LDL-AURIS can therefore represent different information from the Semantic Form Overlap Family Size of the word form.

This chapter does not rule out that the auditory family size effect can be completely understood in terms of discriminative learning in the DLM for auditory word recognition. Future research might derive a measure from the DLM that can account for the full variance explained by family size measures. Such a measure should probably not combine both positive and negative evidence for the association between forms and meanings in a single measure, like Semantic Density does, but purely reflect positive, morphological information.

In conclusion, our results show that the DLM contributes to explaining the variance in the RTs of an auditory lexical decision experiment. Moreover, the DLM can account for parts of the variance that is accounted for by family size measures. Future research has to show what this latter finding means for the DLM: whether a different measure can be derived from DLM implementations that can fully explain family size effects or whether the model first has to be adapted.

Chapter 5

Competing Accounts of the Auditory Family Size Effect: Spreading Activation vs. Discriminative Learning

This chapter is based on:
Müller, H., ten Bosch, L., & Ernestus, M. (2026). Competing accounts of the auditory family size effect: Spreading activation vs. discriminative learning. *Studi AISV.*

Abstract

Words with more morphologically related words are recognised more quickly. We investigated the cognitive mechanisms underlying these family size effects by comparing how well they are accounted for by different models of spoken word recognition. First, we showed that family size effects can partially be accounted for by DIANA, a model of auditory word recognition that does not consider the morphological structures of words. This raises the question of what drives family size effects if it is not morphologically processing. Second, we found that an enriched DIANA, in which we incorporated spreading of activation between morphologically related words, does not better account for auditory family size effects. Third, we showed that discriminative learning accounts somewhat better for family size effects, except for that part that relies on phonological similarity. Together these results suggest that family size effects are more driven by phonological properties of the words than is commonly assumed.

A word's family size (FS) is the number of unique word types that includes the word's root. For example, the word *take* has family members including *takeaway, takes*, *took*, and *uptake*. Research has shown that in both visual (e.g., Moscoso del Prado Martín et al., 2004) and auditory (e.g., Chapter 3) modalities, words with larger morphological families are recognised more quickly (henceforth: the FS effect). This article investigates whether and to what extent the spreading activation mechanism (de Jong, Schreuder, & Baayen, 2003) can account for the FS effect in spoken word recognition. The spreading activation is compared to discriminative learning, which has been shown to partially explain the auditory FS effect (Chapter 4).

This chapter is structured as follows. We first describe the FS effect in more detail and discuss two possible theoretical explanations of this effect. We then detail our research questions. Subsequently, we describe the human spoken word recognition model DIANA (ten Bosch, Boves, & Ernestus, 2022) in more detail and how we adapted this model to investigate our research questions.

5.1 The family size effect

The FS effect has been documented in various studies across multiple languages (e.g., Bertram, Baayen, & Schreuder, 2000; Mulder et al., 2014) for both reading and listening. Researchers propose that this effect is primarily driven by family members that are semantically similar to the word to be recognised (e.g., Moscoso del Prado Martín et al., 2004). Chapter 3 indicates that family members also contribute more to the FS effect the more they are phonologically similar to that word.

There are systematic differences in how family members affect visual versus auditory word recognition. In visual word recognition, the FS effect is independent of the morphological structure of the word to be recognised. That is, the effect has been documented for the recognition of prefixed (e.g., Moscoso del Prado Martín et al., 2004), simplex (e.g., Mulder, Schreuder, & Dijkstra, 2013), and suffixed words (e.g., Bertram, et al., 2000). In contrast, in auditory word recognition, the FS effect seems restricted to simplex and suffixed words (Chapter 3).

This difference between the visual and auditory FS effect can be understood by considering how language users perceive words in the two modalities. Readers

can see most words at once, because most words fit in the readers' foveal area (approximately ten characters; Legge, Mansfield, & Chung, 2001) or in the combination of the foveal area with the parafoveal preview (e.g., Rayner, 1998). This explains why the FS effect surfaces irrespectively of whether the word's root, which represents the commonality among morphological family members, appears as the first, second, or third morphological constituent. In contrast, in auditory word recognition, words unfold over time, and the recognition process begins as soon as the audio signal starts (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Reinisch, Jesse, & McQueen, 2010). While the speech signal unfolds, listeners consider all possible words stored in their memories that match the incoming speech signal perceived so far, and eliminate those that no longer match (e.g., Marslen-Wilson & Welsh, 1978). As a consequence, prefixes are processed first, followed by roots, and then suffixes. This explains why the FS effect has not been attested for spoken prefixed words: the recognition process is well on its way before the word's root is perceived.

5.2 Theoretical explanations for the family size effect

One explanation for the FS effect is based on the theory of discriminative learning (Widrow & Hoff, 1960; Rescorla & Wagner, 1972), as incorporated in the Discriminative Lexicon Model (DLM; e.g., Chuang & Baayen, 2021). This theory assumes that the connections between elements of a word's form (e.g., sequences of graphemes or sounds) and its meaning are strengthened when they co-occur and that they are weakened when they do not. For instance, the connections between sequences of letters within *take* and the meaning of *take* are strengthened by family members of *take* such as *takeaway* and *takes*, which contain the sequence *take* and have a meaning related to that of *take* ('to move something from one place to another'). Weakening occurs if a grapheme or sound sequence (e.g., /bɪə/) appears in words with different meanings (e.g., *beer*, *beard*) or if a morpheme has multiple meanings (e.g., *bank*). Discriminative learning is expected to account for FS effects because the more family members a word has, the stronger the associations become between this word form's features and its meaning, resulting in quicker recognition when the word is encountered. Mulder and colleagues (2014) showed for visual word recognition that (an earlier version of) DLM can indeed account for at least part of the FS effect. They never tested, however, whether the DLM can account for the complete FS effect.

For auditory word recognition, Chapter 4 suggests that the FS effect cannot be completely explained by the DLM. We tested LDL-AURIS (Shafaei-Bajestan et al., 2023), which is an auditory version of the DLM. LDL-AURIS takes as its input a word's audio recordings and it produces as its output a vector representing the meaning of that word. Semantic vectors are numerical representations of words' meanings. Words with more similar meanings tend to have more similar semantic vectors, among other reasons, because semantic vectors reflect how often words co-occur with which other words. We derived from LDL-AURIS the predictor semantic density, which measures the average similarity (in terms of cosine similarity) between an input word's semantic vector and the ten most similar semantic vectors in the mental lexicon. With higher predicted semantic density, a stimulus is more likely to be a word, but also less distinct from other words. In a regression model, this predictor explained part of the same variance of auditory lexical decision response times (RTs) as FS does, suggesting that the FS effect may emerge through discriminative learning.

As an alternative to the DLM, the FS effect may be explained with a spreading activation mechanism. A spreading activation mechanism has been implemented in a few models of visual word recognition, for instance, the Morphological Family Resonance Model (MFRM; de Jong, Schreuder, & Baayen, 2003). The MRFM's lexicon consists of lemma representations (cf. Levelt, 1989), which are connected with syntactic and semantic representations (cf. Schreuder & Baayen, 1995). Semantic representations coincide with morpheme representations, so that family members are linked to the same semantic representation (e.g., *take*, *takeaway*, *takes*, and *uptake* are all connected to the semantic representation *take*), among other semantic representations. When a lemma representation is activated, it transmits activation to all associated syntactic and semantic representations, and those in turn transmit activation to their associated lemma representations, the original lemma representation included. As a consequence, the activation of an encountered word is increased more if it is associated with more words and thus has more family members. As soon as a lemma representation reaches a predefined threshold value for activation, the corresponding word is recognised. The process of propagating activation from lemma representations to central representations and back again is referred to as a cycle, and the number of cycles it needs for a lemma representation to reach threshold activation level determines how quickly a stimulus is responded to.

De Jong et al. (2003) implemented the MFRM to generate lexical decision RTs for visual lexical decision stimuli. They found that the stimuli's Family Sizes correlated with the generated RTs approximately as well as with observed RTs from a lexical decision experiment. This suggests that the MFRM can (at least partly) account for the visual FS effect. Because it has never been investigated whether similar results can be obtained for auditory lexical decision, it is an open question whether the MFRM can account for the auditory FS effect.

5.3 The present chapter: research questions

The present chapter addresses the general question of which cognitive processes underly the auditory FS effect, focusing on the explanatory power of a spreading activation mechanism in comparison to discriminative learning. Because there is no model of human auditory word recognition available that contains a spreading activation mechanism, we had to create such a model. For doing so, we incorporated a spreading activation mechanism into the process-oriented model DIANA (ten Bosch, Boves, & Ernestus, 2022). We addressed the following research questions.

1. To what extent does DIANA without spreading activation incorporated account for the FS effect?
2. Does augmenting DIANA with a spreading activation mechanism enable DIANA to better account for the auditory FS effect?
3. How well does DIANA account for the FS effect compared to LDL-AURIS?

5.4 DIANA

DIANA is a recent model of human auditory word recognition (see ten Bosch, Boves, & Ernestus, 2022, for an overview). In what follows, we describe the original DIANA, which does not contain a spreading activation mechanism. In Chapter 5.4.2, we describe how we incorporated a spreading activation mechanism into DIANA.

DIANA's input is an audio recording, and its output is an ordered list of word hypotheses. These hypotheses and their activation values can be used to predict lexical decision RTs or latencies and amplitudes of EEG components (Bentum et al., 2019). DIANA comprises three main components. The first

component, the activation component, calculates activations or probabilities of both word and pseudoword hypotheses at 10 ms time steps during the unfolding of the audio signal (see Figure 5.1, white blocks). For doing so, the activation component computes feature vectors that resemble Spectro-Temporal Receptive Fields, which simulate phone activations in the auditory cortex (Mesgarani et al., 2014). For each time step (t) between signal onset (t = 0) and signal offset, each word hypothesis' probability is computed as a function of the feature vectors computed from 0 to t:

$$P(\text{hypothesis} \mid \text{feature vectors}_{0:t}) \tag{5.1}$$

This approach mirrors the dynamic nature of human auditory word recognition, which is assumed to start as soon as the speech signal is detected.

The second component, the decision component, evaluates the updated word hypotheses generated by the activation component at every time step to determine the winning (pseudo)word hypothesis, and to conclude whether the audio signal represents a word or a pseudoword. The time required for making the decision between word and pseudoword (lexical decision RT) depends on the dynamic competition between word and pseudoword hypotheses.

Ten Bosch, Boves, and Ernestus (2022) showed that human lexical decision RTs can be well predicted with the decision component considering the entropies of the (pseudo)word hypotheses at word offset, reflecting the principle that resolving ambiguity takes time. Entropy measures the degree of disorder among the hypotheses. The decision component assigns shorter RTs to stimuli with lower entropies at stimulus offset (intercept is a value depending on the experiment, the factor β is a scaling factor in the Hick–Hyman law; Proctor & Schneider, 2018):

$$\text{DecisionTime}_{\text{stimulus entropy}} = \text{intercept} + \beta \sum_i (-p_i \log p_i) \tag{5.2}$$

DIANA's third component, the execution component, mimics neural travelling time, reflecting the time it takes for neural activations in the brain to produce a measurable behaviour. In lexical decision experiments, this behaviour usually is a button press. In DIANA, the execution time is assumed to be constant (i.e., stimulus independent, participant independent).

DIANA's total RT measured from a stimulus' offset depends on the time taken by the decision ('decision time') and the execution ('execution time') as well as on the stimulus duration. Because the longer the stimulus, the more of the processing will have taken place before the end of the stimulus, the factor '*correction offset*' is typically negative:

$$\text{DIANA}_{\text{RT offset}} = \text{correction offset} \cdot \text{stimulus duration} + \text{decision time} + \text{execution time} \quad (5.3)$$

DIANA neither has knowledge about words' morphological structures, nor does it contain spreading activation mechanisms. In addition, setting up DIANA does not involve establishing a mechanism that may implicitly provide information about words' morphological structures, such as form–meaning correspondences in the DLM. We therefore hypothesise that DIANA cannot account for the FS effect. In fact, it may be assumed that words with more family members induce a higher entropy at word offset because the larger a word's FS is, the more probable more family members are, given the audio signal. Consequently, DIANA may be expected to predict longer rather than shorter RTs for larger FS words.

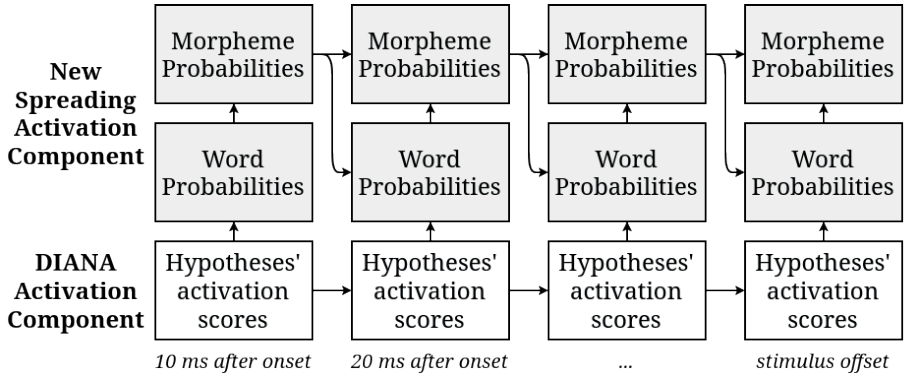


Figure 5.1 Flowchart illustrating how word and morpheme probabilities are updated at each 10ms timestamp in DIANA. White blocks represent DIANA's original activation component. Grey blocks represent the spreading activation mechanism that we incorporated.

5.4.1 Lexicality score

As mentioned above, the original version of DIANA computes an RT basing the decision time on the stimulus entropy (Hick–Hyman law, see fequation 5.2). This implementation implies that the RT is independent of whether the hypothesis that is selected is mostly followed in probability by words or pseudoword

hypotheses. This independence is not entirely plausible for lexical decision tasks, since it can be assumed that lexical decision RTs are (also) based on how likely all real word hypotheses are compared to all pseudoword hypotheses.

We therefore propose an alternative: a decision time based on the *lexicality score*, that is, on a score reflecting the ratio between the probabilities of all real word hypotheses and the probabilities of all pseudoword hypotheses at stimulus offset. The higher the probabilities of the real word hypotheses are in comparison to the probabilities of the pseudoword hypotheses, the shorter the decision time is for a real word. This alternative is again related to entropy, but now applied on the group of words and pseudo words in the candidate hypothesis list, instead of on *individual* hypotheses as was assumed in the 2022-version of DIANA.

The decision time based on lexicality score is the sum of three components (see Equation 5.4). The first component is a constant (intercept). The second component is the word/pseudoword ratio multiplied by a negative scaling factor *lex2RT*. This component reflects the facilitation of a "Yes" response for stimuli with high lexicality scores. The third component is the inverse of this ratio multiplied by a negative scaling factor *invlex2RT*. This component reflects the facilitation of a "No" response for stimuli with low lexicality scores.

$$\text{DecisionTime}_{\text{lexicality score}} = \text{intercept} + \text{lex2RT} \cdot \text{word ratio} + \text{invlex2RT} \cdot \left(\frac{1}{\text{word ratio}}\right) \quad (5.4)$$

In this formula, lex2RT and invlex2RT are the parameters (weight coefficients) that determine to what extent the word ratio and its inverse influence the decision time due to DIANA's decision component.

5.4.2 Augmenting DIANA with a spreading activation mechanism

We implemented into DIANA a spreading activation mechanism that is inspired by the MFRM (de Jong, Schreuder, & Baayen, 2003). We incorporated representations for root morphemes in DIANA's lexicon by adding a new morpheme layer, and we set up connections between word representations and the corresponding morpheme representations. The representations of morphologically complex word hypotheses containing more than one morpheme are connected to multiple morpheme representations (e.g., *background* is connected to the morpheme representations *back* and *ground*). Morpheme probabilities are updated every time the connected words' probabilities are updated (i.e., every 10 ms). The morpheme probabilities, in their turn,

codetermine the probabilities of morphologically related word hypotheses (see Figure 5.1, grey blocks). For more detailed information, see Appendix F.

The spreading activation mechanism as implemented in DIANA involves two meta parameters, *sensitivity* and *decay*, both ranging between 0 and 1. The sensitivity parameter represents the idea that a listener may be more sensitive to the acoustic signal under certain listening conditions (e.g., when hearing with headphones) than under other conditions (e.g., being present at a cocktail party). Mathematically, sensitivity determines the extent to which the probabilities of hypotheses are not only determined by their match with the audio (with a sensitivity value of 1, the probabilities are completely determined by the audio) but also by support from the morphological layer (i.e., top–down expectations). For instance, given a sensitivity of 0.9, if the word *background* has a probability of 0.2 and the morphemes *back* and *ground* have a probability of 0.1 and 0, respectively, the probability of the word *background* will become $0.1 \cdot (0.9 \cdot 0.2 + (1 - 0.9) \cdot (0.1 + 0))$. The decay parameter represents the idea that morphological information slowly disappears when it is not activated anymore. Mathematically, decay determines how much the probability of a morpheme at a time step is determined by its probability at the previous time step and by the probabilities of the connected word hypotheses (with a *decay* value of 1, morpheme probabilities are completely determined by the connected word probabilities of the same time step). For instance, given a decay of 0.5, if the word *running* has a probability of 0.1 and the morpheme *run* has a probability of 0, the probability of the morpheme *run* will become $0.05 \cdot (0.1 \cdot 0.5 + 0 \cdot (1 - 0.5))$.

5.5 General methods

We investigated how well different models (the original DIANA with its decision time based on either stimulus entropy or lexicality score, the spreading activation enriched DIANA, and LDL-AURIS) account for the FS effect on the basis of the Biggest Auditory Lexical Decision Experiment Yet (BALDEY; Ernestus & Cutler 2015, see Chapter 5.5.1). To see whether the different models may take into account the contribution of different types of family members, we tested three different family size measures (see Chapter 5.5.2).

We followed the statistical procedure outlined in Figure 5.2. We first computed a baseline statistical model predicting the RTs from control variables known to predict auditory lexical decision RTs (see Chapter 5.6.1 and see "Baseline"

in Figure 5.2). We then enriched the baseline statistical model such that we produced two enriched statistical models for every word recognition model under study: one enriched statistical model also contained a predictor derived from the word recognition model under scrutiny such as lexicality score or semantic density ("Predictor" in Figure 5.2), while the other enriched statistical model also contained this same predictor in addition to a family size measure ("FS" in Figure 5.2; in interaction with morphological structure, because FS interacts with morphological structure; see Chapter 3). If the latter enriched statistical model better fits the RTs than the former enriched statistical model, the predictor derived from the word recognition model under study does not fully explain the FS effect (see Figure 5.2, third and fourth bar).

In that case, we investigated whether the model of word recognition under study accounts for at least part of the FS effect. We compared how much the FS measure (in interaction with morphological structure) improves the model fit when added to the baseline statistical model (see Figure 5.2, ΔA) and when added to this baseline model enriched with the predictor from the model under study (see Figure 5.2, ΔB). If the FS measure leads to a larger model improvement fit for the baseline model, the model of word recognition at least partially accounts for the FS effect.

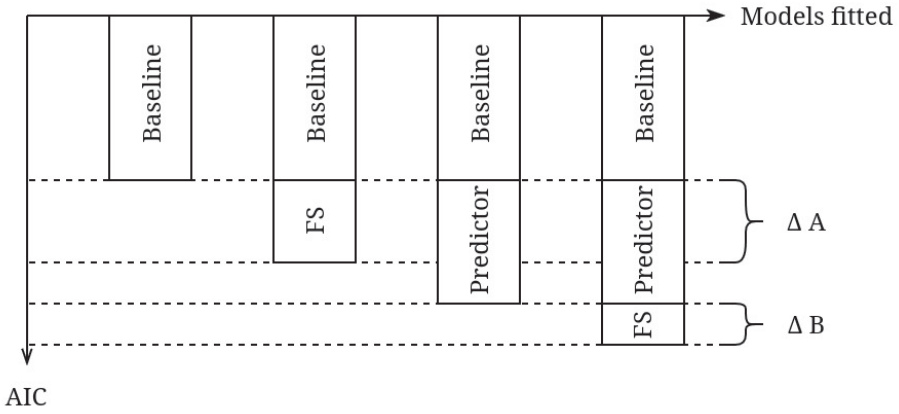


Figure 5.2 Schematic illustration of model comparisons that were carried out to investigate whether a model explains portions of the FS effect. This is the case if adding FS to the baseline model yields a greater model improvement in terms of AIC (ΔA) than adding FS to the baseline model plus the predictor from the model of word recognition under study (ΔB).

The data and the scripts that were used for this chapter can be downloaded from: https://doi.org/10.34973/71n3-vk61.

5.5.1 Data

We analysed the RTs from BALDEY, a Dutch large-scale auditory lexical decision experiment. BALDEY contains RTs from 20 native Dutch speakers on 2,780 spoken Dutch content words and 2,761 pseudowords, both bare roots and inflected and derived words. We only considered accurate responses to real words, excluding compounds, words without morphological analysis in CELEX (Baayen, Piepenbrock, & Gulikers, 1996), and words that do not occur in the Spoken Dutch Corpus (Oostdijk, 2002) and that were therefore not included in training of LDL-AURIS (see Chapter 5.8). Further, we excluded 272 responses (1.71%) that were given before stimulus offset because participants may not have correctly recognised these words, and 24 (0.15%) responses with an encoding error. The resulting dataset contains 15,936 responses in total: 5,908 responses to 322 unique simplex words, 227 responses to 12 unique prefixed words, 8,875 responses to 478 unique suffixed words, and 926 responses to 50 unique words that contain both a prefix and a suffix.

5.5.2 Three family size measures

We focused on three different FS measures that we have tested in Chapter 3. We computed all these measures on the basis of the CELEX database (Baayen, Piepenbrock, & Gulikers, 1996).

Our first FS measure is the word's Classical FS, which is the number of all words with the presented word's root (i.e., all family members). Each family member thus equally contributes to the FS effect.

Our second FS measure is the Semantic FS, which is based on the assumption that a family member contributes the more to the FS effect, the closer its semantic relationship is with the presented word. We quantified the semantic relationship between the presented word and a family member with the help of a distributional semantics model, word2vec (Mikolov et al., 2013), which was trained on more than 600 million messages on Dutch social media, news, blogs, and forums (Nieuwenhuijse, 2018). We computed the cosine of their semantic vectors and normalised the result with min–max scaling so that opposite vectors have a similarity value of 0 and identical vectors have a similarity value of 1. We summed the resulting numbers of all family members. Thus, each family member contributes to the Semantic Family Size according to its semantic similarity with the presented word.

Finally, our third FS measure, Semantic Form Overlap FS, reflects the assumption that family members' contributions to the FS effect are larger with both stronger semantic relationships and larger form overlap with the presented word (Winther Balling & Baayen, 2012). Semantic relationship was calculated on the basis of a distributional semantics model, just as for Semantic FS. Form overlap was based on the words' phonemic representations. That is, the Levenshtein distance between the presented word and a family member was determined, ignoring stress and suffixes; the outcome was squared (emphasising differences in the Levenshtein distance) and inverted (so that greater form overlap results in higher membership values), and passed to the exponential function, resulting in a value between 0 and 1 (i.e., the same range as of semantic relationship values). For each family member, this value was multiplied with its semantic similarity with the presented word (as computed for the Semantic FS) and the outcomes for all family members were summed. This FS measure best predicted RTs for lexical decision in Chapter 3.

Noteworthy, the Classical FS has a perfectly linear relationship with words' numbers of family members. In contrast, the Semantic FS and Semantic Form Overlap FS do not necessarily show linear relationships. For instance, a word with five family members can have a Semantic FS of almost five, when all five family members have strong semantic relationships with this word, but it can also have a Semantic FS of 1, if its family members have weak semantic relationships with this word.

The Classical FS had a mean of 29.69 (SD = 49.83), the Semantic FS a mean of 18.07 (SD = 27.74), and the Semantic Form Overlap FS a mean of 6.65 (SD = 12.96). For the statistical analyses, the measures were log-transformed and subsequently normalised with a z-transformation. The untransformed Classical FS and Semantic FS strongly correlate (r = .995). The untransformed Classical FS and Semantic Form Overlap FS correlate less strongly (r = .782). The same holds for the untransformed Semantic FS and Semantic Form Overlap FS (r = .782).

5.5.3 Control variables in the baseline model

Our baseline statistical model incorporated four control predictors. The first one, Moving Average Response Time (maRT), reflects a participant's local speed and is the weighted average RT across the 10 previous trials (ten Bosch, Ernestus, & Boves, 2018). The second control predictor, trial number (Trial), accounts for a participant's adaptation over the course of the whole experimental session (e.g., Ernestus & Cutler, 2015). The third

control predictor, form frequency (Frequency), captures effects of frequency of occurrence and is derived from CELEX (Baayen, Piepenbrock, & Gulikers, 1996). These three control variables were log-transformed and z-transformed to ensure proper scaling and centring.

Our fourth, categorical, control predictor was morphological structure (MorphStr), which categorises words as "prefixed", "simplex", "suffixed", or "double-affixed" (for words containing both a prefix and a suffix). With this fourth predictor, we can take into account (by means of interactions with the FS measures) that the auditory FS effect differs based on the word's morphological structure (e.g., simplex and suffixed words elicit FS effects whereas prefixed words do not; Chapter 3).

Table 5.1 *Correlation coefficients of the control predictors and FS measures.*

	maRT	Trial	Classial FS	Semantic FS	Semantic Form Overlap FS
Freq	.000	.006	.382	.384	.333
maRT	–	-.003	.005	.005	.005
Trial		–	−.026	−.024	−.019
Classical FS			–	.998	.804
Semantic FS				–	.810

The correlation coefficients between each pair of transformed continuous control variables, as well as between the transformed control variables and either transformed FS measure ranged between .01 and −.03. (see Table 5.1). The exception was Frequency, which showed a weak correlation with each of the three FS measures (range: 0.33–0.38), which is considered unproblematic for the statistical modelling.

5.5.4 General description of the analysis

All statistical models predicted RTs from word offset. They were implemented as Generalised Additive Mixed Models (GAMMs) in R, version 4.0.5 (R Core Team, 2021), with the *mgcv* package (Wood, 2015). GAMMs were selected over Linear Mixed-Effects Models (e.g., Bates et al., 2015) due to their ability to accommodate both linear and non-linear effects. According to Baayen and colleagues (2017), GAMMs yield a better fit to RTs in BALDEY than Linear Mixed-Effects Models.

Our model fitting followed the approach of Bates et al. (2015). That is, the initial baseline model included a by-participant intercept, by-participant random slopes for all continuous variables, thin plate regression splines for continuous control variables, and a parametric term for the categorical variable morphological structure. In subsequent steps, we simplified the model by removing predictors that were not statistically significant. We checked for concurvity, which occurs when one predictor's effect is largely explained by another predictor. If concurvity was above 0.7, we removed the predictor with the smaller p-value, provided it impacted the significance or shape of the remaining predictor's effect.

Eventually, we compared models using standard methods. We evaluated nested models with the χ^2 test on likelihood scores (e.g., Chuang et al., 2021). In case of non-nested models, we considered one model as more likely than another model at the α level of 5% if the former model's Akaike Information Criterion (AIC; Akaike, 1978) was at least 5.88 points (Wagenmaker & Farrell, 2004) smaller than that of the latter model.

5.6 Experiment 1

Experiment 1 addressed our first research question, that is, how well DIANA without spreading activation accounts for the FS effect. After presenting the baseline model, we compared which of the measures that can be extracted from DIANA predicts most reliably the lexical decision RTs: decision time based on stimulus entropy or on lexicality score.

5.6.1 Baseline model

The maximal baseline model that converged is summarised in Table 5.2. It includes a parametric effect for morphological structure, splines for all other control variables, a by-participant intercept and a by-participant random slope for trial number.

The parametric effect for morphological structure is statistically significant for all models tested in the present study. According to the baseline model, simplex words are responded to more slowly than suffixed words ($t = -9.318$, $p < .001$), double-affixed words ($t = -2.921$, $p < .005$), and prefixed words ($t = -2.177$, $p = < .05$). There are no significant differences in RTs between suffixed and prefixed words ($t = 0.143$, $p = .886$), between suffixed and double-

affixed words (t = 1.588, p = .112), or between prefixed and double-affixed words (t = 0.607, p = .544).

Table 5.2 *Summary of the baseline model fitted to log(RT from word offset) in BALDEY. For parametric coefficients, the intercept represents morphologically simple words.*

A. Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	6.263	0.0357	0.036	< .001
MorphStr:Prefixed	−0.091	0.04167	−2.177	0.030
MorphStr:Suffixed	−0.097	0.0104	−9.318	< .001
MorphStr:DoubleAffixed	−0.063	0.022	−2.921	0.004
B. Smooth terms	**Edf**	**Ref.df**	**F**	***p*-value**
s(Frequency)	5.659	6.906	7.704	< .001
s(maRT)	4.798	6.006	289.347	< .001
s(Trial)	5.933	7.105	4.517	< .001
s(Participant)	18.525	19.000	26.586	< .001
s(Trial, Participant)	15.813	19.000	8.911	< .001

The other control predictors' effects are as expected, which shows the reliability of our analyses. RTs were longer the lower the participant's local speed, the further the participant was in the experiment (probably as the result of fatigue), and the lower the frequency of the word.

5.6.2 Stimulus entropy and lexicality score

We first established with which parameter values DIANA best predicts the lexical decision RTs, both when decision time is based on stimulus entropy and when it is based on lexicality score. We set the execution time to 200 ms (a constant assumed to be independent of stimulus and participant). With a gradient search, we found the following optimal parameters: an intercept of 550 ms, a β of 14, an InvLex2RT of −6.1 and an Lex2RT of −85, and, finally, a correction offset of −0.61.

With these optimal values of DIANA's parameters, we determined whether the model better predicts lexical decision RTs when the decision time is based on stimulus entropy or on lexicality score. Both measures are theoretically based on the probabilities of all word and pseudoword hypotheses. We only took into account the 100 most probable hypotheses at word offset because the other hypotheses hardly affect the decision time, either based on stimulus entropy or lexicality score.

Adding either decision time based on stimulus entropy or lexicality score as a predictor to the baseline model produces a significantly better model ($\chi2(2) = 1254.409$, $p < .001$; $\chi2(2) = 2547.34$, $p < .001$, respectively). In terms of AIC, the baseline model enriched with decision time based on stimulus entropy (AIC = 26280.6) fits the data worse than the baseline model enriched with decision time based on lexicality score (AIC = 26252.7). The difference in AIC of 26.96 suggests that the latter model is to be preferred.

5.6.3 How well does DIANA, without spreading activation, account for the family size effect?

To determine whether DIANA, without spreading activation, completely accounts for FS effects, we determined whether a significantly better statistical model is obtained when any FS measure, in interaction with morphological structure, is added as predictor to the baseline model enriched with DIANA's decision time, based either on stimulus entropy or lexicality score (cf. Figure 5.2). Table 5.3 shows that this is the case. This shows that DIANA does not completely account for the FS effects.

In order to see whether DIANA accounts for at least some part of the FS effects, we compared how much the addition of any FS measure improves the baseline model, in terms of AIC, compared to how much it improves the baseline model with DIANA's decision time based on either stimulus entropy or lexicality score as a predictor. The comparison shows that the addition of an FS measure improves the former model more than the latter model (see Figure 5.3). This shows that DIANA accounts for part of the FS effects, even without spreading activation incorporated.

Table 5.3 *Comparison of the performance of the baseline model enriched with a DIANA measure, either based on stimulus entropy or lexicality score, with the performance of the same model further enriched with any of the three FS measures, in interaction with morphological structure.* Sem. FO FS *refers to* Semantic Form Overlap FS.

DIANA measure: decision time based on	Predictor	$\chi2(8)$	*p*-value	AIC Difference
Stimulus Entropy	Classical FS	15.862	< .001	−31.25
Stimulus Entropy	Semantic FS	17.067	< .001	−34.58
Stimulus Entropy	Sem. FO FS	27.308	< .001	−46.44
Lexicality Score	Classical FS	10.992	< .05	−22.35
Lexicality Score	Semantic FS	11.6597	< .005	−24.26
Lexicality Score	Sem. FO FS	20.413	< .001	−31.57

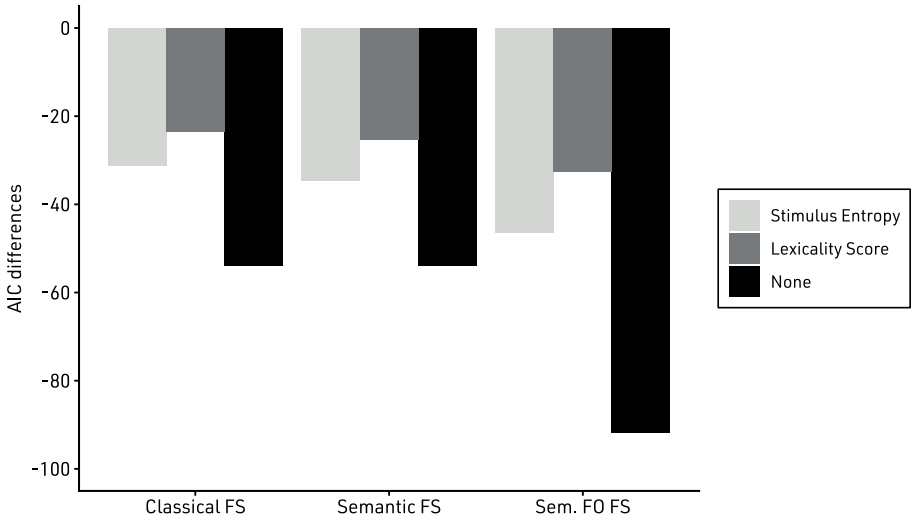


Figure 5.3 Improvement in AIC points when either FS measure is added to the baseline model and to the baseline model with DIANA's decision time based either on stimulus entropy or Lexicality Score. Sem. FO FS refers to Semantic Form Overlap FS.

The last column of Table 5.3 suggests that adding either FS measure results in a smaller model improvement, in terms of AIC when, in the statistical model, DIANA is represented by decision time based on lexicality score than on stimulus entropy. This shows that the measure based on lexicality score better accounts for FS effects than the one based on stimulus entropy. For this reason, Experiments 2 and 3 will only consider decision times based on lexicality score.

5.7 Experiment 2

Experiment 2 investigated our second research question, that is, whether a spreading activation mechanism enables DIANA to more accurately predict RTs and to better account for the FS effects.

5.7.1 Methods

We enriched DIANA with spreading activation as described in Chapter 5.4.2 and in Appendix F. We derived the root morphemes and the morpheme–word relationships from CELEX (Baayen, Piepenbrock, & Gulikers, 1996), which is the largest database for Dutch words with morphological annotations known to us. We obtained the optimal values for the parameters sensitivity and decay

by determining with a grid search with which values DIANA with spreading activation mechanism best predicts the FS effects in our data.

5.7.2 Results

The optimal values of sensitivity and decay depend on whether the spreading activation mechanism is optimised for accounting for Classical FS and Semantic FS (sensitivity = 0.05, decay = 0.825) versus Semantic Form Overlap FS (sensitivity = 0.75, decay = 0.9). While these optimal values hardly differ for decay, they greatly do for sensitivity (0.05 versus 0.75).

The statistical baseline model with DIANA's decision time (based on lexicality score) as additional predictor is equally improved with the addition of an FS measure, in interaction with morphological structure, when DIANA does versus does not incorporate spreading activation (see Figure 5.4). Numerically, incorporating spreading activation in DIANA even leads to a larger contribution of Classical FS (ΔAIC = 0.9) and Semantic FS (ΔAIC = 0.5) to explaining the variance in the data, and hardly any difference in the contribution of Semantic Form Overlap FS (ΔAIC = −0.1). As discussed in Chapter 5.5.4, these differences are statistically insignificant.

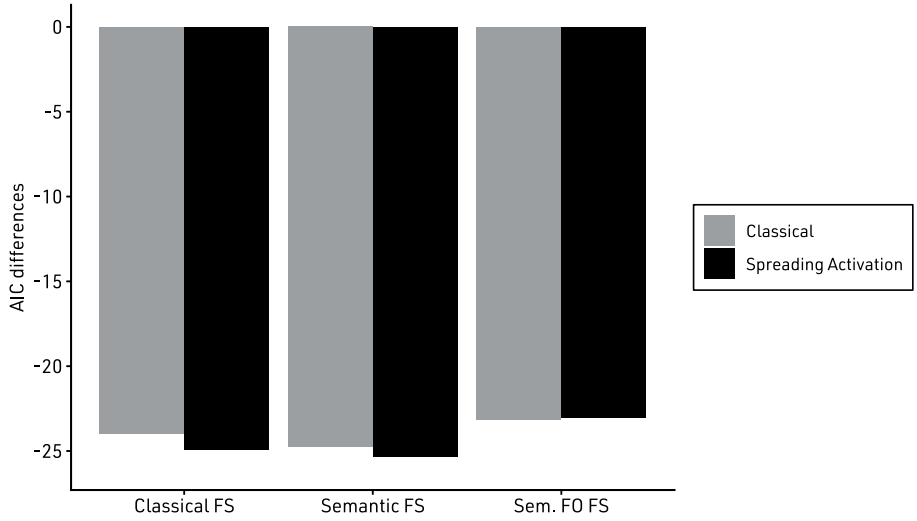


Figure 5.4 AIC improvement (y-axis) yielded by adding one of the three FS measures to either the baseline model enriched with lexicality score from the Classical DIANA or DIANA with spreading activation. Sem. FO FS refers to Semantic Form Overlap FS.

5.8 Experiment 3

Experiment 3 investigated our third research question, that is, how well DIANA accounts for the FS effect relative to LDL-AURIS. Experiment 1 already showed how much the predictor lexicality score derived from DIANA accounts for the FS effects. For determining how much LDL-AURIS accounts for the FS effects in the same data set, we first had to train an LDL-AURIS model, so that we could derive LDL-AURIS' prediction for each word in the dataset, semantic density, from this model.

We trained LDL-AURIS as in Chapter 4. We used *julia* (Bezanson et al., 2017) and the package *JudiLing* (Luo, Chuang, & Baayen, 2020), setting all parameters exactly as Shafaei-Bajestan and colleagues (2023) did. The input training data consisted of the word tokens from Component O (read-aloud speech recordings from Dutch native speakers) of the Spoken Dutch Corpus (Oostdijk, 2000). Because read-aloud speech is typically clearly pronounced, the recordings in Component O are similar to the stimuli of BALDEY. We sliced out word tokens from their acoustic context, drawing on the segmentations of the Spoken Dutch Corpus. We removed mispronounced, incomplete, and unintelligible word tokens, resulting in a dataset of 550,688 word tokens (39,278 word types). As output training data, we provided LDL-AURIS with each word token's semantic vector. Semantic vectors were taken from the Dutch distributional semantics model that we also used to determine Semantics FS and Semantic Form Overlap FS.

5.8.1 Results

Adding the predictor semantic density from LDL-AURIS, in interaction with morphological structure, to the baseline model produces a significantly better model ($\chi 2(8) = 1253.630$, $p < .001$). However, in terms of AIC, the baseline model enriched with semantic density (AIC = 26297.3) fits the data worse than the baseline model enriched with lexicality score (AIC = 26252.7, see Experiment 1). The difference in AIC of 44.6 suggests that DIANA predicts the RTs better than LDL-AURIS does.

The baseline model with semantic density as additional predictor is improved by the addition of any FS measure, in interaction with morphological structure (see Table 5.4). This result is in line with the results from Chapter 4, showing that, like DIANA, LDL-AURIS does not completely account for FS effects.

The improvement in AIC resulting from either FS measure is smaller when this predictor is added – in interaction with morphological structure – to a statistical model with semantic density as a predictor than to a statistical model without this predictor, as can be seen in Figure 5.5. This result is also in line with the result in Chapter 4, showing that LDL-AURIS accounts for parts of the FS effects.

Table 5.4 *Comparison of, on the one hand, the baseline model enriched with semantic density and, on the other hand, the same model further enriched with either of the three FS measures (in interaction with morphological structure).* Sem. FO FS *refers to* Semantic Form Overlap FS

Predictor	$\chi 2(8)$	*p*-value	AIC Difference
FS	23.958	< .05	−10.552
Semantic FS	27.398	< .05	−16.414
Semantic FO FS	50.260	< .001	−29.606

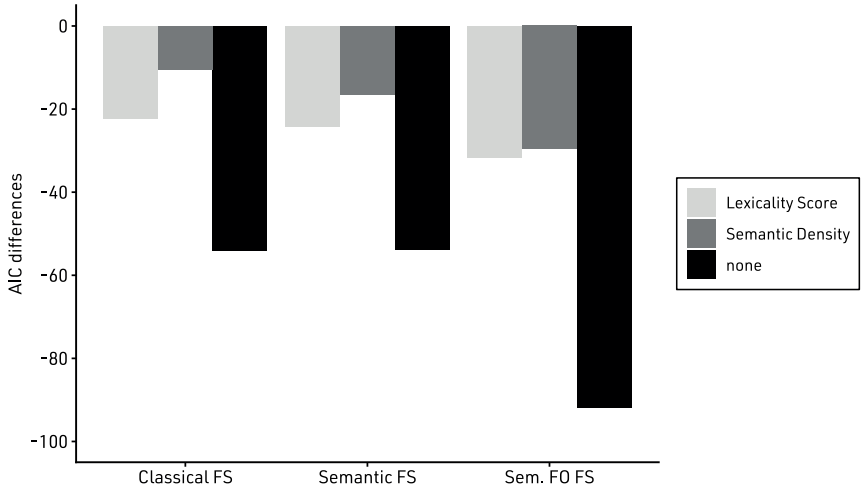


Figure 5.5 Improvement in AIC points (y-axis) when adding either FS measure to the baseline model or the baseline model with lexicality score (see Experiment 1) or semantic density (in interaction with morphological structure). Sem. FO FS refers to Semantic Form Overlap FS.

Importantly, the differences in AIC provided in Tables 3 and 4 (last columns) suggest that adding any FS measure to the baseline model results in a smaller model improvement in terms of AIC when the baseline model already contains semantic density (instead of lexicality score). This difference is significant for Classical FS (ΔAIC = 11.796) and Semantic FS (ΔAIC = 7.844), but not for Semantic Form Overlap FS (ΔAIC = 1.961). LDL-AURIS thus seems to better explain Classical FS and Semantic FS than DIANA (without spreading activation).

5.9 Discussion

This chapter addressed the question whether the auditory family size (FS) effect can be accounted for by the mechanism of spreading activation, which spreads activation from activated words to their morphologically related words, via shared morphemes, and back (de Jong, Schreuder, & Baayen, 2003). We tested the spreading activation mechanism by incorporating it in the spoken human word recognition model DIANA (ten Bosch, Boves, & Ernestus, 2022). We first investigated to what extent DIANA without spreading activation accounts for FS effects. We compared the performance of DIANA's original measure for decision time based on stimulus entropy with the performance of a new one based on lexicality score. Subsequently, we investigated whether DIANA augmented with a spreading activation mechanism better accounts for FS effects. Third, we investigated how well DIANA accounts for the FS effect in comparison to the Discriminative Lexicon Model (DLM; e.g., Chuang & Baayen, 2021). We based our research on the auditory lexical decision response times (RTs) from BALDEY (Ernestus & Cutler, 2015).

5.9.1 DIANA partly accounts for the family size effect

Our results indicate that DIANA without spreading activation accounts for parts of the FS effects. This result is unexpected because this original version of DIANA does not incorporate any morphological knowledge. This finding suggests that the FS effect is not solely driven by the processing of morphological information.

One possible explanation for this unexpected finding is that morphologically related words are typically phonologically similar, and phonologically similar words may affect each other's recognition. However, a higher number of phonologically related words (or continuation forms) in the lexicon typically leads to more uncertainty in the system about which word was exactly uttered and therefore (according to Hick's law) to longer RTs. In contrast, a higher number of morphologically related words leads to shorter RTs.

Another possible explanation is that family size is correlated with word frequency (e.g., Baayen, Tweedie, & Schreuder, 2002), which would be a proxy for syntactic and morphological co-occurrence probabilities (McDonald & Shillcock, 2001) and these syntactic and morphological co-occurrence probabilities are incorporated into DIANA, by means of a language model. Note that, although DIANA can incorporate frequency information (with frequency weightings of lexical hypotheses; ten Bosch, Boves, & Ernestus, 2022), we refrained from this

possibility. We took frequency effects into account by including frequency as a control variable in the regression models.

In sum, this chapter suggests that FS effects are not just the result of morphological processing. Future research is needed that reveals exactly why a model of speech processing that does not involve morphological processing nevertheless accounts for part of the FS effects.

We tested how well DIANA accounts for FS effects by studying two measures that can be derived from DIANA: decision time based on stimulus entropy, which was used in previous studies (see ten Bosch, Boves, & Ernestus, 2022) and decision time based on lexicality score, which we defined in the present chapter and which may be argued to better reflect the processes underlying lexical decision. While both measures predict the auditory lexical decision RTs analysed in the present chapter to some extent and while both partly explain the FS effect, the decision time based on lexicality score performs better in both respects. We therefore addressed the remaining research questions for DIANA's decision time based on lexicality score.

5.9.2 Insights into theoretical accounts of the auditory FS effect

As mentioned above, we tested whether spreading activation between words and morphemes can account for the complete FS effects. We implemented in DIANA morpheme representations and a mechanism that enables word representations to activate morpheme representations and vice versa. This version of DIANA did not account for a larger part of the FS effects than the original version of DIANA. This suggests that a spreading mechanism does not help to explain the FS effect.

Chapter 4 suggests that discriminative learning can at least partly account for auditory FS effects. We investigated whether discriminative learning, as implemented in LDL-AURIS (Shafaei-Bajestan et al., 2023) better accounts for the auditory FS effect than DIANA does, by testing the two models on the basis of the same data set. Interestingly, we found that LDL-AURIS yields a worse fit to our RTs than DIANA, but better accounts for the Classical FS effect and Semantic FS effect than DIANA. One explanation for this finding is that LDL-AURIS is based on a theory about how morphological information is processed, which is informed by a great body of research (e.g., Baayen et al., 2011; Baayen & Smolka, 2020; Tomaschek et al., 2021), while the development of DIANA is based on abundant research on speech processing.

DIANA and LDL-AURIS account for the effect of the Semantic Form Overlap FS to the same extent. One explanation for why DIANA does not perform worse – as it does for the effects of the Classical FS and the Semantic FS – may be that Semantic Form Overlap FS explicitly takes into account how well each word stored in the lexicon matches the presented audio signal, which forms the core activity of DIANA. Given that DIANA does not incorporate morphological processing, and given that Semantic Form Overlap FS best accounts for lexical decision RTs (Chapter 3), this result suggests that family size effects may be more driven by phonological properties of the words than is commonly assumed.

5.9.3 The role of alternative model implementations

The results obtained in the present chapter may raise the question to what extent they depend on how exactly we formulated the spreading activation mechanism in DIANA. For theoretical reasons, our implementation of the spreading activation mechanism deviated in two aspects from the implementation of the spreading activation mechanism in the MFRM by de Jong, Schreuder and Baayen (2003). First, while de Jong and colleagues implemented a spreading activation mechanism between lemma representations on the one hand and representations for semantics, syntactical classes, and affix representations on the other hand, for theoretical reasons, we implemented in DIANA a spreading activation mechanism between word and morpheme representations. Second, we did not incorporate cycles to compute response times. Alternative implementations of the spreading activation mechanisms may better account for auditory FS effects.

When implementing the spreading activation mechanism, we realised that it is unclear how exactly activation should be propagated from words to morphemes and vice versa. In our implementation (see Appendix F), we assumed that a word transmits its probability equally to all its morphemes. For instance, if the word *butterfly* has a probability of 0.5, the morphemes *butter* and *fly* each have a probability of 0.5 too (that is, if their probabilities are not co-determined by their probabilities at the previous time step). In contrast, we assumed that how strongly a morpheme's probability affects a word's probability depends on the word's number of morphemes: the probability propagated by a morpheme is its probability divided by the number of morphemes in the word. For instance, if the morpheme *butter* has a probability of 0.6, the word *butter* receives a probability of 0.6 from the spreading activation mechanism, whereas the word *butterfly*, which has two morphemes, only receives half as much support. There are obviously alternative assumptions possible, and unfortunately, it is

difficult to favour one assumption above another based on existing theories. Future research could investigate whether a spreading activation mechanism based on (slightly) different assumptions better accounts for FS effects.

Finally, our implementation of the spreading activation mechanism takes into account the relevance of form overlap between the presented word and the family member for the FS effect (e.g., Winther Balling & Baayen, 2012), by means of the sensitivity parameter: the higher this parameter, the less family members that do not substantially match the audio signal contribute to spreading of activation. In contrast, this implementation does not take into account the relevance of semantic similarity between the presented word and the family member for the FS effect (e.g., Moscoso del Prado Martín et al., 2004). Future research could try to bring semantic similarity into the spreading activation mechanism.

5.10 Conclusion

The present chapter shows that auditory FS effects are partly explained by the original version of DIANA, which does not include morphological processing. This raises questions about the exact nature of FS effects. Augmenting DIANA with connections between word and morpheme representations does not enable DIANA to explain a greater portion of the family size effects and decreases how accurately it accounts for lexical decision RTs. A spreading activation mechanism, at least as we implemented it, thus cannot account for auditory FS effects. The principle of discriminative learning as implemented in LDL-AURIS accounts for larger parts of the FS effects. However, when the family size effects are based both on semantic and form similarity, the two models perform equally well. Together this suggests that explaining the auditory family size effect may require the combinations of two theories: a theory of how morphological information is processed and a theory of audio signal processing.

Chapter 6

General Discussion and Conclusions

This thesis examined three main research questions related to the role of morphology in spoken word recognition. Concisely formulated, these questions were the following: First, does the morphological structure of a spoken word influence how listeners recognise this word? Second, which mechanisms underly effects of morphological structure in spoken word recognition? Third, which insights can be gained by investigating the recognition of spoken morphologically complex words with the research method of computational modelling. To answer these questions, I carried out four empirical and computational studies. These studies were presented in Chapters 2–5. In this sixth, final chapter, I will discuss which conclusions can be drawn from these studies with respect to the above-mentioned questions.

6.1 Effects of morphology in human spoken word recognition

The studies presented in this thesis provide converging evidence that human spoken word recognition is influenced by words' morphological structures. In Chapter 2, I investigated whether the recognition of morphologically complex spoken words is governed by their roots, their surface forms, or a combination of both. If root information contributes to recognition, either on its own or alongside surface forms, this would suggest that listeners unconsciously draw on knowledge of morphological structure during the word recognition process. I analysed six response time (RT) datasets from two large-scale lexical decision experiments, employing three conceptually different, statistical models. The three statistical models represent three systematically different theoretical assumptions with respect to the word recognition process. These assumptions are the following. First, word recognition is root-driven; second, word recognition is driven by surface forms; and third, recognition is governed by whichever frequency dominates, root or form frequency, with a clear tipping point marking the switch between the two. With these three models, I analysed RTs to Dutch suffixed words that are formed with either the suffixes *-en*, *-t*, or *-heid*, and the RTs were either obtained in visual or in auditory lexical decision experiments.

The analysis of the RTs to spoken words with the suffix *-en* did not provide conclusive evidence about whether any of the three above-mentioned statistical models more accurately predicts the RTs than any of the remaining models. In contrast, for written words with the suffix *-en*, the RTs were most

accurately predicted with the tipping point model. The RT distributions for the other four datasets – responses to written or spoken words with the suffixes *-t* or *-heid* – were most accurately predicted with the form-driven model. To validate these results, I have analysed synthesised data to test whether it is possible to determine which model is most likely to have generated a given RT distribution.

I assume that recognising spoken words involves one single cognitive architecture, independent of the word structure. Based on this assumption, the results from Chapter 2 suggest that the recognition of morphologically complex words is both, driven by their roots and their forms: forms exert a greater influence for the studied stimuli (four of six analysed datasets), but roots play a role too (one of six datasets). This leads to the question of when recognition is root- and when it is form-driven. Given the results from Chapter 2, the tipping point between root- and form-driven processing cannot be (solely) based on the ratio between root and form frequency. Chapter 2 provides weak cues that root-driven processing is more likely for a) written word recognition in contrast to spoken word recognition, and b) for more productive than less productive affixes.

Another insight from Chapter 2 is that the tipping point is not a categorical threshold but rather a gradual transition. For written nouns with the suffix *-en*, I found that recognition shifts progressively as the balance between root and form frequency changes. Plurals with root frequencies being at least ten times higher than the corresponding form frequencies were mostly recognised on the basis of their roots, while plurals with root frequencies being at most three times higher than the corresponding form frequencies were mostly recognised on the basis of their surface forms. The recognition of plurals with root frequencies at least ten times greater than their form frequencies was mostly root-driven, whereas the recognition of plurals with root frequencies no more than three times greater were mostly form-driven. Crucially, plurals with intermediate ratios (root frequencies three to ten times greater than form frequencies) exhibited both root- and form-driven processing. This pattern suggests that root-driven and form-driven recognition should be viewed as two ends of a continuum, with a transitional range where both mechanisms contribute.

Taken together, these findings from Chapter 2 indicate that morphological structure plays a systematic role in shaping word recognition, but not in a

categorical way. To further explore how morphological structure affects recognition beyond root and form frequencies, Chapters 3–5 focus on another key phenomenon: the morphological family size effect. The family size effect supposes that the more words can be derived from a word's root, the faster a word with this root is on average responded to in lexical decision experiments. While the family size effect was often replicated in visual lexical decision experiments (e.g., Schreuder & Baayen, 1997; de Jong, Schreuder, & Baayen, 2000; Juhasz & Berkowitz, 2011), previous research on the family size effect in auditory lexical decision experiments has produced inconclusive findings (Baayen, Wurm, & Aycock, 2007; Winther Balling & Baayen, 2008, 2012; Wurm et al., 2006). In Chapter 3, based on the analysis of RTs from one large-scale visual and one large-scale auditory lexical decision experiment in Dutch, I found that the family size also affects word recognition in the auditory modality. The finding that not only written words but also spoken words elicit the facilitative morphological family size effect represents a new insight. This finding indicates, in addition to the findings from Chapter 2, that morphological structure affects how spoken morphologically complex words are recognised. Because this finding is based on the analysis of the largest number of RTs (20,493 observations) that were analysed in order to investigate the auditory family size effect yet, it is the most reliable finding in the literature so far regarding the auditory family size effect.

The results of Chapter 3 suggest that previous studies were not able to consistently obtain the facilitative auditory family size effect because they did not systematically take into account the interaction between words' morphological structures and the family size effect. That is, the results of Chapter 3 suggest that spoken simplex and spoken suffixed words show the family size effect, whereas words containing a prefix do not. That prefixes block the family size effect is most likely due to the fact that, in spoken word recognition, the root, as the most informative cue to a morphological family, is perceived only after the prefix and thus the processing of the root begins most likely after the processing of the prefix. In written word recognition, in contrast, the presence of a prefix may not delay the perception of the word's root to the same extent, because the prefix and the root can be simultaneously perceived. This is the case because many words wholly fit into the central vision span, which covers approximately ten characters (Legge, Mansfield, & Chung, 2001), and words that do not wholly fit into this span are likely perceived as a whole due to parafoveal preview (e.g., Rayner, 1998).

To find out which mechanisms drive the family size effect, I tested six different variants to measure family size, whereby each measure represents different assumptions with respect to whether the family size effect is driven by semantic, form, or both semantic and form relationships between the family members. The measure that most accurately predicted the RTs represents the assumption that morphological family members contribute stronger to the family size effect, the greater their semantic similarity and form overlap with the presented word is, suggesting that the family size effect is driven by both semantic similarities and form overlap. The finding that the family size effect is driven by semantic similarities is consistent with findings on the visual family size effect (e.g., Bertram, Baayen, & Schreuder, 2000; Moscoso del Prado Martín et al., 2004; Schreuder & Baayen, 1997). A new insight that is generated by Chapter 3 is that semantic similarities can adequately be quantified with a Distributional Semantics Model. Another new insight gained is the fact that the family size effect is also driven by the form overlap between family members and the presented word, which seems to contradict previous studies on the visual family size effect. These studies claim that form overlap does not modulate the family size effect (Bertram, Baayen, & Schreuder, 2000; de Jong, 2002). However, these studies have used a strict dichotomous conceptualisation of form overlap, differentiating between family members that either completely or incompletely overlap with the presented word. In contrast, I have used a gradient measure, gauging the degree to which family members overlap in form with the presented word.

6.2 How morphological structure affects a morphologically complex word's recognition

Considering the indications for the effects of morphological structure on word recognition presented in Chapters 2 and 3, the question arises as to which theories can best explain these effects. I investigated two different theories of word recognition by implementing these theories as computational models and then testing how accurately these models can predict the morphological family size effect in the auditory lexical decision dataset that I analysed in Chapter 3. The first theory is of a distributional–connectionist nature. It eschews the concept of morphemes (Chapter 4). The second theory is localist in the sense that it supposes that the mental lexicon consists of explicit representations of linguistic units such as words and morphemes (Chapter 5).

In Chapter 4, as an implementation of distributional–connectionist theory, I investigated whether and to what extent the Discriminative Lexicon Model (DLM; e.g., Chuang & Baayen, 2021) accounts for the auditory family size effect. The DLM suggests that family size effects emerge because morphological family members have both some form and some meaning in common, which is, from a theoretical perspective, in line with the findings from Chapter 3. According to the DLM, these form and meaning commonalities among morphological family members help to structure the mapping of a word form onto a meaning without requiring explicit morpheme representations; the more family members a word has, the easier it is to map this word's form onto its corresponding meaning, resulting in a faster lexical decision.

Chapter 4 represents the first study that applies the DLM implementation LDL-AURIS (Shafaei-Bajestan et al., 2023) to predict auditory lexical decision RTs. For predicting the RTs, I derived two predictors from LDL-AURIS that have also been derived from other DLM implementations to predict visual lexical decision RTs (Heitmeier, Chuang, & Baayen, 2023; Heitmeier et al., 2024). The first predictor measures how strongly the presented word's meaning, as gauged by a Distributional Semantics Model (e.g., Bruni, Tran, & Baroni, 2004), correlates with the meaning predicted by LDL-AURIS. The second predictor measures the average distance of all words' meanings contained in the Distributional Semantics Model to the meaning predicted by LDL-AURIS. Contrary to expectations, the first predictor did not predict the RTs that I analysed and therefore was not further considered. The second predictor yielded a facilitative instead of the previously observed (Heitmeier, Chuang, & Baayen, 2023) inhibitory effect. From these findings can be concluded that the DLM as implemented in LDL-AURIS positively contributes to the prediction of auditory lexical decision RTs, but we do not understand how these predictions come about, because the two established predictors elicit different effects than previously reported. It is a common problem that connectionist–distributional models function as a black box that may yield relatively accurate predictions without providing insights into the processes that lead to these predictions (e.g., Guidotti et al., 2018).

Furthermore, the results from Chapter 4 suggest that LDL-AURIS can predict a part of the family size effect. It only partly explains the family size effect, because after explaining as much variance in the RTs as possible with the help of control variables and LDL-AURIS, there is still variance in the RTs left over, which can be accounted for by the presented words' family size. This finding

suggests that parts of the family size effect can be explained without assuming that the mental lexicon comprises representations for morphemes. However, to put these findings into perspective, it is essential to also take into account the findings from Chapter 5.

In Chapter 5, I investigated whether and to what extent the human spoken word recognition model DIANA (ten Bosch, Boves, & Ernestus, 2022) can account for the auditory family size effect (in comparison to LDL-AURIS). DIANA is a process-oriented end–to–end model that takes an audio signal as input and produces word hypotheses as output. I tested two different approaches for predicting lexical decision RTs with DIANA. First, I investigated how well the entropy of DIANA-generated hypotheses predict the RTs. Entropy measures a signal's degree of informativeness. Stimuli that result in many hypotheses with more similar probabilities are less informative than stimuli that result in one hypothesis with a clearly high probability. I found that lower entropy stimuli and thus more informative stimuli elicit faster RTs than higher entropy stimuli. Second, I investigated how well the ratio between the likelihood of lexical hypotheses on the one hand and on the other hand the likelihood of non-lexical hypotheses predicts the lexical decision RTs. I called this predictor the *lexicality score*. I found that the more likely a stimulus is a lexical item according to DIANA (i.e., higher lexicality scores), the faster the stimulus is recognised. Lexicality score appeared to be a better predictor of lexical decision RTs than entropy, suggesting that lexical decisions may not be the result of successful word recognition but rather successful recognition of lexicality (how similar the stimulus is to known words).

For testing whether and to what extent DIANA explains the family size effect, I investigated both stimulus entropy and lexicality score. I found that both stimulus entropy and lexicality score partly account for the family size effect and that lexicality score accounts for a greater portion of the family size effect. However, after explaining as much variance in the RTs as possible with the help of control variables and either stimulus entropy or lexicality score, there is still variance in the RTs left over, which can be accounted for by the presented words' family size.

That stimulus entropy and lexicality score partially explain the family size effect is unexpected because both predictors ignore any information about morphological relationships. I speculate that the following mechanism explains why the predictors explain parts of the family size effect: the more

family members the presented word has, the more likely family members exist that have a large form overlap with the audio signal. Due to the large form overlap, these family members represent probable hypotheses, resulting in low entropies. In addition, these high probability hypotheses will shift the ratio of words to pseudowords (within the 100 most probable hypotheses, say) in favour of words, resulting in higher lexicality scores. Thus, larger family sizes lead to lower entropies and higher lexicality scores, which predict faster RTs. Future research should thoroughly investigate this hypothetical explanation for why stimulus entropy and lexicality score can partly account for the family size effect.

In Chapter 5, I also investigated whether and to what extent the family size effect can be accounted for by a spreading activation mechanism similar to the mechanism described in the Morphological Family Resonance Model (MFRM; de Jong, Schreuder, & Baayen, 2003). This spreading activation mechanism follows the tradition of localist theories, in so far as it draws on concrete morpheme and word representations that become activated in the course of the word recognition process, because morphologically related words activate each other via shared morpheme representations. According to this model, the family size effect emerges because the more family members a word has, the more activation it receives back from the shared morpheme representations, resulting in a quicker recognition. A previous study has shown that RTs produced by the MFRM correlate with family size, suggesting that a spreading activation mechanism potentially provides an explanatory account for the visual family size (de Jong Schreuder, & Baayen, 2003). Because RTs produced by the MFRM do not completely coincide with family size, it remains unclear whether the MFRM wholly accounts for the family size effect or only to some extent. Because of this, I investigated whether augmenting DIANA with a spreading activation mechanism similar to the MFRM enables the augmented DIANA to account for a greater portion of the family size effect than DIANA without a spreading activation mechanism.

For implementing a spreading activation mechanism in DIANA, I augmented DIANA's lexicon with morpheme representations and connections between morpheme and word representations via which activation can be propagated between the two types of representations. From this augmented DIANA, I derived new lexicality scores. In Chapter 5, I show that augmenting DIANA does not enable lexicality score to account for the family size effect to a greater extent than when lexicality score is derived from DIANA without an

augmentation, suggesting that the auditory family size effect does not come about as a consequence of spreading activation.

A comparison of how well lexicality scores from DIANA and semantic density scores from the DLM account for the auditory family size effect showed that both models' predictors perform similarly when family members are weighted by their form and semantic similarity to the target word (cf. Chapter 3). On this basis, no firm conclusion can be drawn as to whether family size effects arise from a discriminative learning mechanism or whether they are instead an epiphenomenon of how probability distributions over possible words are structured.

6.3 Using computational modelling to study human spoken word recognition

For answering this thesis' research questions, the technique of computational modelling has proven itself as a powerful toolkit. With computational modelling techniques it is possible to compare non-standard inferential statistical models that represent different assumptions about the role of morphological processing mechanisms in word recognition (Chapter 2), to shed light on the driving forces of the family size effect (Chapter 3, 4, and 5), and to compare distributional–connectionist theories with localist theories (Chapter 5). Computational modelling helps to explore to what extent the family size effect can be understood in light of discriminative learning (Chapter 4) or spreading activation (Chapter 5), which was a previously missing, albeit crucial piece of information in current research on morphological processing. Furthermore, with the aid of computational modelling, I was able to provide substantial evidence that the accuracy with which the human spoken word recognition model DIANA (e.g., ten Bosch, Boves, & Ernestus, 2022) predicts lexical decision RTs is significantly higher than the accuracy with which LDL-AURIS (Shafaei-Bajestan et al., 2023) predicts these RTs (Chapter 5), which is a landmark insight for future developments of human spoken word recognition models. For the investigation of the above-mentioned matters, different computational approaches were utilised such as statistical implementations of theoretical models (Chapter 2), data synthesis techniques (Chapter 2), (nested) model comparisons (Chapters 2, 3, 4, and 5), and the evaluation of end-to-end models (Chapters 4 and 5), showcasing that computational modelling techniques can aid in the exploration of a variety of

theories and models and are therefore of paramount importance in the study of theories of human word recognition

6.3.1 The importance of well-defined theoretical concepts

Although the method of computational modelling is suitable for answering a variety of research questions, this method has its limitations and poses challenges, which ideally should be addressed by future research. One of these challenges is that statistical and computational models can only be based on theoretical concepts that are well-defined. Ambiguous concepts often become apparent as such only when one tries to implement a given theory as a statistical or computational model. For instance, the statistical models in Chapter 2 assume that the duration of morphologically complex words' decomposition depends on these words' cumulative stem frequencies. I had difficulties applying the definition of cumulative stem frequency provided by Baayen, Dijkstra, and Schreuder (1997) to words of other word classes than nouns, because their cumulative stem frequency is based on a distinction between inflectional and derivational affixes. However, such a distinction is not clearly defined. The alternative concept of root frequency (e.g., Taft, 1979, 2004) is also founded on a distinction between inflectional and derivational affixes. Because of this, I invented the bi-morphemic root frequency, which is the summed frequency of all words that share a root and that consist of at most two morphemes. This definition is unambiguous and can thus be straightforwardly computed for words with different word classes.

Different definitions of frequency of occurrence (e.g., cumulative stem frequency, root frequency, form frequency) tap into different assumptions about the word recognition process. For instance, if the contribution of a morphologically complex word's root to this word's recognition depends on frequencies of words that are derived from this word, but not on this word's inflected forms, this implies that derived and inflected forms do not interact with each other in the course of word recognition. If root frequency but not base frequency determines decomposition times, this implies that the recognition of morphologically complex words involves access to these words roots and not their bases. Thus, the problem of finding a sound definition of cumulative stem frequency demonstrates the importance of formulating our theories about the word recognition process in a precise and unambiguous manner.

In Chapter 3, the challenging task of deciding on a definition of a theoretical concept reoccurred in the context of the form overlap between morphologically

related words. The notion of form overlap represents the assumption that words' matches with the auditory signal modulate how strongly these words contribute to the family size effect. I quantified form overlap based on the Levenshtein distance between phonemic representations, ignoring stress and suffixes. This quantification of form overlap may be too simplistic. From a theoretical point of view, phonetic features (e.g., Albright, 2007) provide a more fine-grained representation of words' form and therefore may be a more useful basis for quantifying form overlap. Future research should also consider including stress in form overlap. For instance, the Dutch words *doorlopen* /ˌdoːrˈloːpə(n)/ 'to go through' and *doorlopen* /ˌdoːrˈloːpə(n)/ 'to keep going' contribute to the same extent to the family size of *doorlooppuntjes* /ˈdoːrˌloːpˈpʏnˌtjəs/ 'suspension points', although the items differ in stress pattern. Because there are some indications that the duration of a root vowel depends on whether the root is followed by another segment (Kemps et al., 2005; Salverda, Dahan, & McQueen, 2003), future research should also investigate whether fine phonetic detail affects how strongly morphological family members contribute to the family size effect.

6.3.2 The difficulty of operationalising theoretical concepts

Even if theoretical concepts are well-defined, it may be challenging to validly and reliably operationalise these concepts. To illustrate this, consider the concept of semantic similarity. Semantic similarity is a well-defined measure, especially in the context of a Distributional Semantics Model, but different Distributional Semantics Models may yield different semantic similarities. In Chapter 3, 4, and 5, I based semantic similarity on a Distributional Semantics Model that was trained with *word2vec* (Mikolov et al., 2013). One reason for choosing a word2vec-based model instead of, for instance, a *fasttext*-based model (Bojanowski et al., 2017), is that the word2vec algorithm does not operate on sub-word information such as n-grams, which tend to converge with morphemes. Because fasttext-based semantic similarities may capture properties of morphological family sizes, weighing family members' contribution to the family size effect with fasttext-based semantic similarities may result in a weaker family size effect, because fasttext already consists of information about morphological relationships. Thus, the number of morphological family members may already be entailed in the fasttext vector, which would then account for the same variance in the RTs as family size. I also could have used a Distributional Semantics Model that is partially grounded in visual images (e.g., Pennington, Socher, & Manning, 2014; Shahmohammadi, Lensch, & Baayen, 2021). From a theoretical perspective, it is difficult to gauge

how such a Distributional Semantics Model relates to a word2vec-based model and thus it becomes unclear how semantic measures based on these two models relate. When measurements from different Distributional Semantics Models may yield different predictions, more research is needed to better understand how these differences come about, especially as the different models increasingly often are used in linguistic research (e.g., Heitmeier, Chuang, and Baayen, 2023).

6.3.3 What we can (and cannot) learn from lexical decision experiments

This thesis is based on the analysis of lexical decision RTs. It is important to note that these RTs are the outcome of a temporal process. It is assumed that as soon as the auditory signal begins to unfold (e.g., Marslen-Wilson & Welsh, 1978), participants start to process it, but they will usually make their decisions only after word offset (Ernestus & Cutler, 2015). Usually, these RTs are analysed with statistical regression models that estimate the size of different effects but that do not reflect the temporal order of these effects. For instance, there are indications that the frequency effect is especially strong early on in the recognition process (Dahan, Magnuson, & Tanenhaus, 2001; Dufour, Brunellière, & Frauenfelder, 2013). Such temporal dynamics of word recognition cannot be captured with regression models. Therefore, future analyses of lexical decision RTs should also be conducted with other statistical models that provide insights into the temporal order of predictors' effects. Such models could be statistical models such as piece-wise exponential additive mixed models (Hendrix & Sun, 2019) or computational models such as DIANA (ten Bosch, Boves, & Ernestus, 2022; Chapter 5).

It is important to keep in mind that, while analysing lexical decision RTs with other techniques may shed more light on the temporal dynamics of the word recognition process, the gain from these techniques will be limited, because lexical decision RTs only represent the final outcome of the temporal process; what exactly happens before the lexical decision is made and in which order processes such as lexical or semantic access happen, is difficult to gauge on the basis of the RTs. Because of this, it is necessary to investigate morphological processing in the context of other experimental data than lexical decision RTs too. A promising method for investigating how listeners process spoken words and speech is the analysis of EEG-data. For instance, Bentum et al. (2022) presented participants with naturalistic speech. For each word, they computed two word probability distributions. The first probability distribution was based on the word's preceding context and the second probability distribution was an update of the

first probability distribution that was computed by evaluating the word's audio information with DIANA's phone and word activations. They found that greater differences between the first and the second probability distribution inversely correlated with the phonological mismatch negativity, indicating that listeners build expectations about what they are going to hear, and these expectations are constantly updated based on what they hear.

Lexical decision data do not only provide restricted insights into temporal processes underlying word recognition, but they also obscure our idea of what word recognition is. That is, to make a correct response to a word stimulus, it would be sufficient for the participant to recognise any word. For instance, when presented with *decisions*, participants can easily overhear the final *-s* and correctly make a "yes-response", because *decision* is a word. The fact that a predictor based on the ratio between lexical and non-lexical hypotheses among the best hypotheses significantly predicts lexical decision RTs (Chapter 5) suggests that listeners indeed do not necessarily recognise the presented word but just evaluate the stimulus as any word. I would speculate that semantic processing in lexical decision is rather shallow and thus more similar to the shallow semantic processing of word naming (e.g., Balota et al., 2004) than to the semantic processing of a task such as comprehension listening, which requires to integrate a word's meaning into a discourse model (Martin, 2021). If semantic processing in lexical decision is rather shallow, measures and models that operate on semantic representations such as the Semantic Family Size (Chapter 2) or the DLM (Chapter 4) may more accurately model the processing of naturalistic speech than lexical decision RTs. However, I do not know of any studies that investigate family size measures or the DLM in the context of naturalistic speech processing.

Finally, the lexical decision task may affect the *mode of processing*. That is, participants encounter plenty of morphologically complex words that they usually do not encounter in isolation. For instance, while it is reasonable to assume that citation forms such as infinitives (e.g., *maken* 'to make') do occur in isolation with some regularity in everyday language, it is highly unlikely that past participle verbs occur in isolation in everyday language (e.g., *gemaakt* 'made' as in 'he has made'). Because of this artificial nature of lexical decision experiments, participants may have been more sensitive to morphological structures during the experiment, which may have increased the role of morphological processing and thus the morphological family size effect (greater awareness of morphological family members).

6.4 Conclusion

This thesis sheds light on the question of how morphological structure affects the recognition of spoken words. I found that the recognition of spoken morphologically complex words can be driven by their morphological constituents and that the likelihood with which these constituents drive the recognition may depend on the word's suffix (Chapter 2). I found that the number of a word's morphological family members affects how quickly this word is recognised and that the influence exerted by the family members depends on their form overlap and semantic similarity with the presented word. In addition, I found that the auditory family size effect interacts with the presented word's morphological structure (Chapter 3). Another insight won is that LDL-AURIS can predict auditory lexical decision RTs and that it partly explains the family size effect (Chapter 4). However, DIANA also explains aspects of the family size without relying on knowledge about words' morphological makeups (Chapter 5). Contrary to expectations, augmenting DIANA with information about words' morphological relationships did not improve its ability to explain the family size effect. These results suggest that effects often attributed to morpheme-based representations – such as the morphological family size effect – can emerge without assuming mental representations of morphemes. In conclusion, using various computational modelling techniques, I gathered empirical support for the role of morphological structure in spoken word recognition and demonstrated that computational models of human spoken word recognition need to capture this role in order to more accurately describe the recognition process.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40*(1), 41–61. https://doi.org/10.1006/jmla.1998.2607

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Amenta, S., & Crepaldi, D. (2012). Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in Psychology, 3,* 232. https://doi.org/10.3389/fpsyg.2012.00232

Amenta, S., Crepaldi, D., & Marelli, M. (2020). Consistency measures individuate dissociating semantic modulations in priming paradigms: A new look on semantics in the processing of (complex) words. *Quarterly Journal of Experimental Psychology, 73*(10), 1546–1563. https://doi.org/10.1177/1747021820927663

Amenta, S., Marelli, M., & Sulpizio, S. (2017). From sound to meaning: Phonology-to-semantics mapping in visual word recognition. *Psychonomic Bulletin & Review, 24*(3), 887–893. https://doi.org/10.3758/s13423-016-1152-0

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, H. R. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLoS ONE, 12*(4), e0174623. https://doi.org/10.1371/journal.pone.0174623

Baayen, R. H., Burani, C., & Schreuder, R. (1997). Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1996* (pp. 13–33). Springer. https://doi.org/10.1007/978-94-017-3718-0

Baayen, R. H., Chuang, Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity, 2019,* 4895891. https://doi.org/10.1155/2019/4895891

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*(1), 94–117. https://doi.org/10.1006/jmla.1997.2509

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*(2), 290–313. https://doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., Lieber, R., & Schreuder, R. (1997). The morphological complexity of simplex nouns. *Linguistics, 35*(5), 861–877. https://doi.org/10.1515/ling.1997.35.5.861

Baayen, R. H., McQueen, J. M., Dijkstra, T., & Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (Trends in Linguistics Studies and Monographs, Vol. 151, pp. 355–390). Mouton de Gruyter.

Baayen, R. H., Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychological Review, 118*(3), 438–482. https://doi.org/10.1037/a0023851

Baayen, R. H., Neijt, A. (1997). Productivity in context: A case study of a Dutch suffix. *Linguistics, 35*(3), 565–588. https://doi.org/10.1515/ling.1997.35.3.565

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX lexical database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., & Linke, M. (2020). An introduction to the generalized additive model. In S. T. Gries & M. Paquot (Eds.), *A practical handbook of corpus linguistics* (pp. 563–591). Springer.

Baayen, R. H., & Smolka, E. (2020). Modeling morphological priming in German with naïve discriminative learning. *Frontiers in Communication, 5*, 17. https://doi.org/10.3389/fcomm.2020.00017

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language, 81*(1–3), 55–65. https://doi.org/10.1006/brln.2001.2506

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon, 2*(3), 419–463. https://doi.org/10.1075/ml.2.3.06baa

Bauer, L., Lieber, R., & Plag, I. (2015). *The Oxford reference guide to English morphology.* Oxford University Press.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. *arXiv Preprint,* arXiv:1506.04967. https://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity.* John Wiley & Sons.

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language, 42*(3), 390–405. https://doi.org/10.1006/jmla.1999.2681

Bertram, R., Laine, M., & Karvinen, K. (1999). The interplay of word formation type, affixal homonymy, and productivity in lexical processing: Evidence from a morphologically rich language. *Journal of Psycholinguistic Research, 28*(3), 213-226. https://doi.org/10.1023/A:1023200313787

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 489–511. https://doi.org/10.1037/0278-7393.26.2.489

Bentum, M., Ten Bosch, L., Van den Bosch, A., & Ernestus, M. (2019, September). Listening with great expectations: An investigation of word form anticipations in naturalistic speech. In *Interspeech 2019: 20th Annual Conference of the International Speech Communication Association* (pp. 2265–2269). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2019-2741

Beyersmann, E., Cavalli, E., Casalis, S., & Colé, P. (2016). Embedded stem priming effects in prefixed and suffixed pseudowords. *Scientific Studies of Reading, 20*(3), 220–230. https://doi.org/10.1080/10888438.2016.1140769

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*(1), 65–98. https://doi.org/10.1137/141000671

Blevins, J. P. (2016). *Word and paradigm morphology.* Oxford University Press.

Boudelaa, S., & Marslen-Wilson, W. D. (2011). Productivity and priming: Morphemic decomposition in Arabic. *Language and Cognitive Processes, 26*(4–6), 624–652. https://doi.org/10.1080/01690965.2010.521022

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology, 28*(5), 803–809. https://doi.org/10.1016/j.cub.2018.01.080

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8*(3), 531–544. https://doi.org/10.3758/BF03196189

Burnham, K. P., & Anderson, D. R. (2004). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition, 28*(3), 297–332. https://doi.org/10.1016/0010-0277(88)90017-0

Cho, J., Pires, A., & Brennan, J. R. (2024). How large are root and affix priming effects in visual word recognition? Estimation from original data and a Bayesian meta-analysis. *Language, Cognition and Neuroscience, 39*(10), 1291–1309. https://doi.org/10.1080/23273798.2024.2384051

Chuang, Y. Y., & Baayen, R. H. (2021). Discriminative learning and the lexicon: NDL and LDL. In Oxford Research Encyclopedia of Linguistics. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.375

Chuang, Y. Y., Fon, J., Papakyritsis, I., & Baayen, R. H. (2021). Analyzing phonetic data with generalized additive mixed models. In M. Ball & N. Müller (Eds.), *Manual of Clinical Phonetics* (pp. 108–138). Routledge.

Chuang, Y. Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2021). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods, 53*(3), 945–976. https://doi.org/10.3758/s13428-020-01356-w

Colé, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language, 28*(1), 1–13. https://doi.org/10.1016/0749-596X(89)90025-9

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology, 47*(2), 109–121. https://doi.org/10.1016/S0022-2496(02)00016-0

Dawson, N., Rastle, K., & Ricketts, J. (2018). Morphological effects in visual word recognition: Children, adolescents, and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(4), 645–654. https://doi.org/10.1037/xlm0000485

De Grauwe, S., Lemhöfer, K., & Schriefers, H. (2019). Processing derived verbs: The role of motor-relatedness and type of morphological priming. *Language, Cognition and Neuroscience, 34*(8), 973–990. https://doi.org/10.1080/23273798.2019.1599129

De Jong, N. H. (2002). *Morphological families in the mental lexicon* (Doctoral dissertation, Radboud University Nijmegen).

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., & Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language, 81*(1–3), 555–567. https://doi.org/10.1006/brln.2001.2547

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes, 15*(4–5), 329–365. https://doi.org/10.1080/01690960050119625

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2003). Morphological resonance in the mental lexicon. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 65–88). Mouton de Gruyter.

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R., & Baayen, R. H. (2005). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes, 20*(1–2), 7–41. https://doi.org/10.1080/01690960444000124

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology, 68*(8), 1469–1488. https://doi.org/10.1080/17470218.2014.984730

Farrar, D. E., & Glauber, R. G. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics, 49*(1), 92–107. https://doi.org/10.2307/1937887

Feldman, L. B., & Siok, W. W. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(3), 776–781. https://doi.org/10.1037/0278-7393.23.3.776

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., & Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods, 50*(4), 1285–1307. https://doi.org/10.3758/s13428-017-0943-1

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*(2), 488–496. https://doi.org/10.3758/BRM.42.2.488

Gabry, J., & Češnovar, R. (2022). *cmdstanr: R interface to 'CmdStan'* (Version 0.5.3) [Computer software]. https://mc-stan.org/cmdstanr/

Geary, J. A., & Ussishkin, A. (2018). Root-letter priming in Maltese visual word recognition. *The Mental Lexicon, 13*(1), 1–25. https://doi.org/10.1075/ml.18001.gea

Gonnerman, L. M., Seidenberg, M. S., & Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General, 136*(2), 323–345. https://doi.org/10.1037/0096-3445.136.2.323

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518–565. https://doi.org/10.1037/0033-295X.103.3.518

Gravel, S., Bigman, J. S., Pardo, S. A., Wong, S., & Dulvy, N. K. (2024). Metabolism, population growth, and the fast–slow life history continuum of marine fishes. *Fish and Fisheries, 25*(2), 349–361. https://doi.org/10.1111/faf.12811

Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior, 2*(1), 1–11. https://doi.org/10.1007/s42113-018-0011-7

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics, 39*(6), 1041–1070. https://doi.org/10.1515/ling.2001.041

Hay, J., & Baayen, R. H. (2002). Parsing and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 2001* (pp. 203–235). Springer. https://doi.org/10.1007/978-94-017-3726-5

Heitmeier, M., Chuang, Y. Y., & Baayen, R. H. (2023). How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology, 146,* 101598. https://doi.org/10.1016/j.cogpsych.2023.101598

Heitmeier, M., Chuang, Y. Y., Axen, S. D., & Baayen, R. H. (2024). Frequency effects in linear discriminative learning. *Frontiers in Human Neuroscience, 17,* 1242720. https://doi.org/10.3389/fnhum.2023.1242720

Jackendoff, R., & Audring, J. (2020). Relational morphology: A cousin of construction grammar. *Frontiers in Psychology, 11,* 2241. https://doi.org/10.3389/fpsyg.2020.02241

Jared, D., & O'Donnell, K. (2017). Skilled adult readers activate the meanings of high-frequency words using phonology: Evidence from eye tracking. *Memory & Cognition, 45*(3), 334–346. https://doi.org/10.3758/s13421-016-0661-4

Juhasz, B. J., & Berkowitz, R. N. (2011). Effects of morphological families on English compound word recognition: A multitask investigation. *Language and Cognitive Processes, 26*(4–6), 653–682. https://doi.org/10.1080/01690965.2010.498668

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1,* 174. https://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 English words. *Behavior Research Methods, 44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*(7–8), 1089–1132. https://doi.org/10.1080/01690960802193688

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: Toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 876–895. https://doi.org/10.1037/a0013484

Legge, G. E., Mansfield, J. S., & Chung, S. T. (2001). Psychophysics of reading: XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision Research, 41*(6), 725–743. https://doi.org/10.1016/S0042-6989(00)00295-9

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press.

Lõo, K., Järvikivi, J., Tomaschek, F., Tucker, B. V., & Baayen, R. H. (2018). Production of Estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Morphology, 28*(1), 71–97. https://doi.org/10.1007/s11525-017-9318-7

Luke, S. G., & Christianson, K. (2011). Stem and whole-word frequency effects in the processing of inflected verbs in and out of a sentence context. *Language and Cognitive Processes, 26*(8), 1173–1192. https://doi.org/10.1080/01690965.2010.510359

Lüdeling, A., & De Jong, N. H. (2002). German particle verbs and word formation. In N. Dehé, R. Jackendoff, A. McIntyre, & S. Urban (Eds.), *Verb-particle explorations* (pp. 315–339). Mouton de Gruyter. https://doi.org/10.1515/9783110902341.315

Luo, X., Chuang, Y. Y., & Baayen, R. H. (2020). JudiLing: An implementation in Julia of Linear Discriminative Learning algorithms for language model. Eberhard Karls Universität Tübingen, Seminar für Sprachwissenschaft.

Magnusson, M., Andersen, M. R., Jonasson, J., & Vehtari, A. (2020). Leave-one-out cross-validation for Bayesian model comparison in large data. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* (Vol. 108, pp. 341–351). PMLR. https://proceedings.mlr.press/v108/magnusson20a.html

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92,* 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods, 50*(4), 1482–1495. https://doi.org/10.3758/s13428-018-1017-8

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of orthography–semantics consistency on word recognition. *Quarterly Journal of Experimental Psychology, 68*(8), 1571–1583. https://doi.org/10.1080/17470218.2014.959709

Marslen-Wilson, W. D. (1980). Speech understanding as a psychological process. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Spoken language generation and understanding* (pp. 39–67). Springer. https://doi.org/10.1007/978-94-009-9091-3_2

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*(3), 295–322. https://doi.org/10.1177/00238309010440030101

Meunier, F., & Segui, J. (1999). Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language, 41*(3), 327–344. https://doi.org/10.1006/jmla.1999.2642

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Milin, P., Smolka, E., & Feldman, L. B. (2018). Models of lexical access and morphological processing. In G. Libben, M. Goral, & G. Libben (Eds.), *The Oxford handbook of the mental lexicon* (pp. 230–255). Oxford University Press. https://doi.org/10.1002/9781118829516.ch11

Miller, B., Juhasz, B. J., & Rayner, K. (2006). The orthographic uniqueness point and eye movements during reading. *British Journal of Psychology, 97*(2), 191–216. https://doi.org/10.1348/000712605X66845

Miller, R. B. (1968). Response time in man-computer conversational transactions. In *Proceedings of the AFIPS '68 Fall Joint Computer Conference (Part I, Vol. 33, pp. 267–277)*. AFIPS. https://doi.org/10.1145/1476589.1476628

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. arXiv. https://arxiv.org/abs/1301.3781

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*(2), 165–178. https://doi.org/10.1037/h0027366

Morton, J. (1970). A functional model of memory. In D. A. Norman (Ed.), *Models of human memory* (pp. 203–254). Academic Press.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1271–1278. https://doi.org/10.1037/0278-7393.30.6.1271

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., & Baayen, R. H. (2005). Changing places: A cross-language perspective on frequency and family size in Dutch and Hebrew. *Journal of Memory and Language, 53*(4), 496–512. https://doi.org/10.1016/j.jml.2005.07.003

Mulder, K., Dijkstra, T., & Baayen, R. H. (2015). Cross-language activation of morphological relatives in cognates: The role of orthographic overlap and task-related processing. *Frontiers in Human Neuroscience, 9,* 16. https://doi.org/10.3389/fnhum.2015.00016

Mulder, K., Dijkstra, T., Schreuder, R., & Baayen, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language, 72,* 59–84. https://doi.org/10.1016/j.jml.2013.12.004

Mulder, K., Schreuder, R., & Dijkstra, T. (2013). Morphological family size effects in L1 and L2 processing: An electrophysiological study. *Language and Cognitive Processes, 28*(7), 1004–1035. https://doi.org/10.1080/01690965.2012.733013

Müller, H., ten Bosch, L., & Ernestus, M. (2024). Can the Discriminative Lexicon Model account for the family size effect in auditory word recognition? *Nota Bene, 1*(2), 176–192. https://doi.org/10.1075/nb.00010.mul

Müller, H., ten Bosch, L., & Ernestus, M. (2024). The family size effect in visual and auditory word recognition. *Language, Cognition and Neuroscience, 39*(6), 793–814. https://doi.org/10.1080/23273798.2024.2337941

Nenadić, F., Tucker, B., & ten Bosch, L. (2023). Computational modeling of an auditory lexical decision experiment using DIANA. *Language and Speech, 66*(3), 564–605. https://doi.org/10.1177/00238309221111752

New, B., Ferrand, I., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*(1), 45–52. https://doi.org/10.3758/BF03193811

Nieuwenhuijse, A. (2018). *Dutch word2vec model* [Data set]. GitHub. https://github.com/coosto/dutch-word-embeddings

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of LREC 2000* (pp. 887–894). ELRA.

Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In P. Peters, P. Collins, & A. Smith (Eds.), *New frontiers of corpus research* (pp. 105–112). Rodopi. https://doi.org/10.1163/9789004334113_008

Pain, M. T. G., & Hibbs, A. (2007). Sprint starts and the minimum auditory reaction time. *Journal of Sports Sciences, 25*(1), 79–86. https://doi.org/10.1080/02640410600718004

Perdijk, K., Schreuder, R., Baayen, R. H., & Verhoeven, L. (2012). Effects of morphological family size for young readers. *British Journal of Developmental Psychology, 30*(3), 432–445. https://doi.org/10.1111/j.2044-835X.2011.02053.x

Perrone-Bertolotti, M., Kujala, J., Vidal, J. R., Hamame, C. M., Ossandon, T., Bertrand, O., & Lachaux, J. P. (2012). How silent is silent reading? Intracerebral evidence for top-down activation of temporal voice areas during reading. *Journal of Neuroscience, 32*(49), 17554–17562. https://doi.org/10.1523/JNEUROSCI.2982-12.2012

Pexman, P. M., & Hargreaves, I. S. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review, 15*(1), 161–167. https://doi.org/10.3758/PBR.15.1.161

Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes, 20*(2), 261–290. https://doi.org/10.1080/01690960444000098

Proctor, R. W., & Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology, 71*(6), 1281–1299. https://doi.org/10.1080/17470218.2017.1322622

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes, 15*(4–5), 507–537. https://doi.org/10.1080/01690960050119689

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Reifegerste, J., Meyer, A. S., & Zwitserlood, P. (2017). Inflectional complexity and experience affect plural processing in younger and older readers of Dutch and German. *Language, Cognition and Neuroscience, 32*(4), 471–487. https://doi.org/10.1080/23273798.2016.1247213

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology, 63*(4), 772–783. https://doi.org/10.1080/17470210903104412

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods, 50*(4), 1568–1580. https://doi.org/10.3758/s13428-017-0981-8

Savinova, E., & Malyutina, S. (2021). Evidence for dual-route morphological processing across the lifespan: Data from Russian noun plurals. *Language, Cognition and Neuroscience, 36*(6), 730–745. https://doi.org/10.1080/23273798.2021.1879182

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 257–294). Lawrence Erlbaum Associates.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*(1), 118–139. https://doi.org/10.1006/jmla.1997.2510

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., & Baayen, R. H. (2021). LDL-AURIS: A computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience. 38*(4), 509–536. https://doi.org/10.1080/23273798.2021.1954207

Solomyak, O., & Marantz, A. (2009). Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. *Brain and Language, 108*(3), 191–196. https://doi.org/10.1016/j.bandl.2008.09.004

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience, 22*(9), 2042–2057. https://doi.org/10.1162/jocn.2009.21296

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv.* https://doi.org/10.48550/arXiv.2008.10296

Stan Development Team. (2022). *Stan modeling language users guide and reference manual* (Version 2.31) [Computer software manual]. https://mc-stan.org

Stevens, P., & Plaut, D. C. (2022). From decomposition to distribution: An attractor-network account of the lexical-semantic system. *Psychonomic Bulletin & Review, 29*(6), 1673–1702. https://doi.org/10.3758/s13423-022-02086-0

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition, 7*(4), 263–272. https://doi.org/10.3758/BF03197599

Taft, M. (2023). Localist lexical representation of polymorphemic words: The AUSTRAL model. In D. Crepaldi (Ed.), *Linguistic morphology in the mind and brain* (pp. 152–166). Routledge. https://doi.org/10.4324/9781003159759-11

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 638–647. https://doi.org/10.1016/S0022-5371(75)80051-X

ten Bosch, L., Boves, L., & Ernestus, M. (2022). DIANA, a process-oriented model of human auditory word recognition. *Brain Sciences, 12*(5), 681. https://doi.org/10.3390/brainsci12050681

ten Bosch, L., Ernestus, M., & Boves, L. (2018). Analyzing reaction time sequences from human participants in auditory experiments. In *Proceedings of Interspeech 2018* (pp. 971–975). ISCA. https://doi.org/10.21437/Interspeech.2018-1728

ten Bosch, L., Boves, L., Tucker, B., & Ernestus, M. (2015). DIANA: Towards computational modeling reaction times in lexical decision in North American English. In *Proceedings of Interspeech 2015* (pp. 1576–1580). ISCA. https://hdl.handle.net/11858/00-001M-0000-0029-1D65-1

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71,* 249–267. https://doi.org/10.1016/j.wocn.2018.09.004

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2021). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics, 57*(1), 123–161. https://doi.org/10.1017/S0022226719000203

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods, 51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Vaknin-Nusbaum, V. (2025). Morphological density and reading comprehension in Hebrew novice readers. *Reading and Writing, 38*(3), 699–721. https://doi.org/10.1007/s11145-024-10526-7

Vaknin-Nusbaum, V., & Shimron, J. (2011). Hebrew plural inflection: Linear processing in a Semitic language. *The Mental Lexicon, 6*(2), 197–244. https://doi.org/10.1075/ml.6.2.01vak

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using GAMMs* (Version 2.3) [R package manual]. https://cran.r-project.org/web/packages/itsadug/

Vehtari, A. (2020). *Cross-validation FAQ.* GitHub repository. https://github.com/stan-dev/loo/blob/HEAD/vignettes/online-only/faq.Rmd

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wagenmakers, E. J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/fα noise in human cognition. *Psychonomic Bulletin & Review, 11*(4), 579–615. https://doi.org/10.3758/BF03196615

Wilder, R. J., Goodwin Davies, A., & Embick, D. (2019). Differences between morphological and repetition priming in auditory lexical decision: Implications for decompositional models. *Cortex, 116,* 122–142. https://doi.org/10.1016/j.cortex.2018.10.007

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. In *1960 WESCON Convention Record* (Part IV).

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70,* 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language, 90*(3), 669–692. https://doi.org/10.1353/lan.2014.0064

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE, 6*(9), e23613. https://doi.org/10.1371/journal.pone.0023613

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics, 59,* 122–143. https://doi.org/10.1016/j.wocn.2016.09.004

Winther Balling, L., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes, 23*(8), 1159–1190. https://doi.org/10.1080/01690960802201010

Winther Balling, L., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition, 125*(1), 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Wood, S. N. (2015). *mgcv: Mixed GAM computation vehicle with automatic smoothness estimation* (Version 1.29) [R package]. https://cran.r-project.org/package=mgcv

Wurm, L. H. (1997). Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language, 37*(3), 438–461. https://doi.org/10.1006/jmla.1997.2524

Wurm, L. H., & Aycock, J. (2003). Recognition of spoken prefixed words: The role of early conditional root uniqueness points. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 259–286). Mouton de Gruyter.

Wurm, L. H., & Ross, S. E. (2001). Conditional root uniqueness points: Psychological validity and perceptual consequences. *Journal of Memory and Language, 45*(1), 39–57. https://doi.org/10.1006/jmla.2000.2758

Wurm, L. H., Ernestus, M. T., Schreuder, R., & Baayen, R. H. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and conditional root uniqueness points. *The Mental Lexicon, 1*(1), 125–146. https://doi.org/10.1075/ml.1.1.08wur

Zou, Y., Tsang, Y. K., Shum, Y. H., & Tse, C. Y. (2023). Full-form vs. combinatorial processing of Chinese compound words: Evidence from mismatch negativity. *International Journal of Psychophysiology, 187,* 11–19. https://doi.org/10.1016/j.ijpsycho.2023.02.004

R

Appendices

Appendix A

We assume that the predicted RT to a simplex word is the sum of the intercept β_0, a set of predictors $x_1,\ldots,x_i$ for I predictors, multiplied with their corresponding betas $\beta_1,\ldots,\beta_i$, and a by-participant intercept u_j for J participants. A word's root and form frequency belong to the set of predictors $x_1,\ldots,x_i$. Simplex words are always recognised through their roots, which are identical to their whole forms, regardless of the processing model. Because RTs to simplex words depend on the root frequency FQ_{root} (Baayen, Dijkstra, & Schreuder, 1997), the RT to a simplex word is predicted by:

$$RT_{\text{Simplex}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \beta_{\text{FQroot}} FQ_{\text{root}} + u_j \quad \text{(A.1)}$$

The root-driven, form-driven, and tipping point model produce different predictions for a complex word's RT. In the root-driven model, recognition of complex words is always root-driven, which implies that plural recognition involves the root frequency plus an additional parsing penalty Δp, resulting in the following predicted response time:

$$RT_{\text{Complex}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \beta_{\text{FQroot}} FQ_{\text{root}} + \Delta p + u_j \quad \text{(A.2)}$$

The form-driven model, in contrast, assumes that recognition of complex words is always form-driven and that the time needed for recognising the whole form is a function of the form frequency FQ_{form}, without parsing penalty. Thus, the predicted RT to a complex word reads:

$$RT_{\text{Complex}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \beta_{\text{FQform}} FQ_{\text{form}} + u_j \quad \text{(A.3)}$$

The tipping point model combines formulae (A.2–A.3). Depending on which processing mode would predict response the quickest, the equation for predicting the RTs to complex words is based on the form frequency (if form-driven recognition is faster) or the root frequency plus an additional parsing penalty (if root-driven processing is faster):

$$RT_{\text{Complex}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \min\left(\beta_{\text{FQform}} FQ_{\text{form}}, \beta_{\text{FQroot}} FQ_{\text{root}} + \Delta p\right) + u_j \quad \text{(A.4)}$$

During model fitting, $x_1,...,x_i$ are provided so that $\beta_0,..., \beta_i$, u_j, and Δp can be estimated. This estimation is comparable to fitting a linear mixed-effects model (cf. Sorensen & Vasishth, 2015), except for the *min*-argument, which is necessary to determine the fastest processing mode.

Appendix B

To prove that our methodological approach can correctly determine whether a distribution of RTs has likely been generated by a root-driven model, a form-driven model, or a tipping point model, we conducted a dedicated validation study. For doing so, we used a two-step procedure. First, we synthesised RTs with three theoretical processing models based on each of the three processing views (i.e., with formulae (B.1–B.3). Second, we investigated whether the theoretical processing model used for synthesising the RTs also most accurately predicts these RTs.

B.1 Data synthesis

We generated RTs from invented, henceforth synthesised, participants reacting to invented, henceforth synthesised, simplex and complex words. For doing so, we assigned properties to the synthesised participants and synthesised words that play a role in the formulae of the three theoretical processing models, that is, root frequency, form frequency and word length/duration (see (B.1–B.3)). We wanted the generated RTs' distributions to closely resemble the RTs' distributions of simplex and suffixed words that are to be analysed in the subsequent experiments, which helps to ensure that the conclusions drawn are valid. To ensure the greatest comparability, we based these properties of our synthesised words (i.e., words' lengths/durations, root frequencies, and form frequencies; x_1, F_{root}, and F_{form} in the formulae) on nouns in BALDEY (Ernestus & Cutler, 2015). In addition, we determined, also on the basis of DLP and BALDEY, what the effects of these properties were on the participants' RTs to the synthesised words. We will now discuss each of these steps in detail.

We generated 100 sets of RTs with each of the three models (B.1–B.3). Each set consists of 6,000 RTs, produced by 20 synthesised participants, responding to 150 synthesised simplex words and 150 synthesised complex words. We based the properties of our synthesised words (i.e., words' durations, root frequencies, and form frequencies; x_1, F_{root}, and F_{form} in the formulae) on nouns in BALDEY (Ernestus & Cutler, 2015). We only analysed the 504 nouns that can form their plural only according to the scheme *root + -s* and that cannot occur as verbs. Table B.1 lists descriptive statistics of the words' form frequencies, root frequencies and durations.

Table B.1 *Means and standard deviations (SD) of the words' durations, root frequencies, and form frequencies, on which we based the properties of our synthesised words in the simulation experiment.*

Variable	Mean	SD
Form frequency	4.118611	1.7497953
Root frequency	5.841478	1.9881844
Duration	6.485120	0.2692574

In order to determine the effect sizes of the variables in our generating models (B.1–B.3), that is, of duration, root frequency, form frequency, trial number (i.e., β_1, β_2, F_{root}, and F_{form}), and by-participant intercepts (u_j), we determined their effect sizes on the RTs for the nouns in BALDEY on the basis of which we had also determined the frequencies and durations of our synthesised words (as listed in Table B.1). We excluded incorrect responses, responses made before word offset, and seven observations made in session 8 by Participant 1 because of erroneous encodings. The resulting constrained dataset comprises 9,742 observations. We fitted a Bayesian linear mixed-effects model to the log-transformed RTs, including form frequency, root frequency, trial number and duration, as predictors, all of which were also log-transformed, as well as by-participant random intercepts. The resulting estimates are presented in Table B.2.

Table B.2 *Estimated effect sizes with corresponding standard deviations (SD) for the effects in BALDEY that are incorporated in the synthesis of RTs in the simulation experiment.*

Predictor	Coefficient	Coefficient SD
Intercept	5.620486325	0.089528972
by-participant effect	–	0.242652274
Duration	0.236941020	0.010009395
Trial	0.009320023	0.002562212
Root frequency	−0.002597006	0.001428608
Form frequency	−0.017916674	0.001626829

We then generated the RTs of a single set as follows. We assumed 300 words with characteristics randomly sampled from the distributions shown in Table B.1 and ascribed a unique trial number in the interval {log(1), log(2), … , log(300)}. For instance, a synthesised word could have a duration of e^6, which would correspond to 403 ms. For determining how much the synthesised duration contributes to the synthesised RT, the duration would be multiplied with 0.237,

which is the effect of duration (see Table B.2). Thus, 96 ms of the synthesised RT are due to the synthesised word's duration. Similarly, we randomly sampled 20 by-participant random intercepts from a normal distribution with Mean = 0 and the estimated SD reported in Table B.2 (0.243). Using the effect sizes' estimates listed in Table B.2, we thus generated 300 RTs for each of the 20 synthesised participants with each of the three theoretical processing models (formulae (B.1–B.3)). We repeated this procedure 100 times, resulting in 100 × 6,000 = 600,000 generated RTs for each model.

Because the estimated root frequency effect is relatively small in comparison to the estimated form frequency effect (Table B.2), our generated RTs to the assumed complex words would never be produced via the root-driven mechanism in our tipping point model. To ensure that the recognition of at least some RTs to complex words were root-driven, we multiplied the effect estimate for root frequency with 50. As a consequence, the tipping point model produced complex words via root-driven mechanisms across the 100 repetitions of the generating procedure between 30% and 49% of the trials.

B.2 Analysis of synthesised data

We fitted each theoretical processing model to the 300 sets of generated RTs to determine its goodness of fit with each set. We provided the model with the generated words' variables and their RTs and let the model estimate the variables' effect sizes and the (by-participant) intercept(s). We used as priors for our models the distributions of the effects that were used for generating the RTs (and which were deduced from the actual data).

B.3 Results

We first investigated whether the sets of RTs generated with the three different models are substantially different. Only if they are different, we can expect that it is possible to determine by which model a set was generated. For each pair of models (root-driven versus form-driven, root-driven versus tipping point, and form-driven versus tipping point), we compared the correlation between each of the 100 RT sets generated by the first model with the corresponding set generated by the other model, which resulted in 100 correlation coefficients for each pair of models. Figure B.1 is divided into three cells, each showing a histogram that represents the 100 correlation coefficients for one pair of models. For example, the histogram in the leftmost cell shows correlation coefficients between the 100 pairs of RT sets generated with the root-driven (100 sets) and the tipping point model (100 sets). All correlation coefficients

range between r_{min} = .92 and r_{max} = .99 and the correlations between RT distributions from the form-driven model and the tipping point model are especially high. The overall strong correlations indicate that the models generate relatively similar RTs, which makes it difficult to tease the generating models apart on the basis of the RT distributions. The RT distributions are so similar because the RTs are mostly determined by the control predictors (e.g., a word's duration) and less so by the processing mode of a complex word (root-driven or form-driven). The RTs from the form-driven and the tipping point model are particularly similar because the tipping point model, when incorporating a relatively high parsing penalty, can behave like a form-driven model.

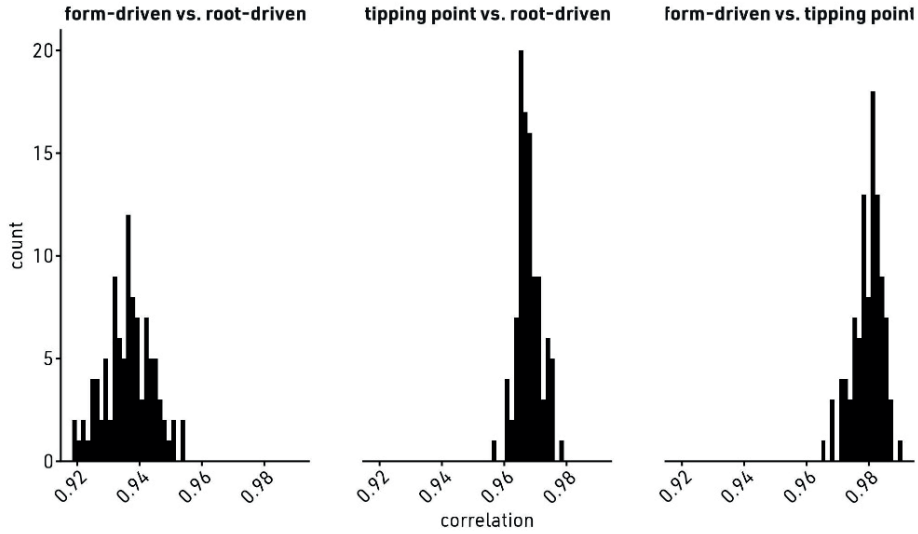


Figure B.1 Histogram of the coefficients of the correlations (x-axis) between the 100 pairs of synthesised response times that were synthesised with the root-driven and the form-driven model (first column), with the root-driven and the tipping point model (second column), and with the form-driven and the tipping point model (third column).

We then investigated whether the original generating models can be determined on the basis of the RT distributions. Figure B.2 shows the difference in ELPD LOO between the root-driven model, the form-driven model, and the tipping point model relative to the model with the highest ELPD LOO, for each set of generated RT distributions. The figure is split up into nine cells, which cross the type of generating model (in rows) and the type of inferential model tested (in columns). Each value in a cell on the x-axis represents one of the 100 generated distributions of RTs produced by the pertinent model. It can

be seen that, although the RT distributions from the three models are highly correlated, the model that generated the RTs makes the most accurate predictions of the RT distributions. The exception is formed by the RTs generated by the form-driven model, which are as accurately fitted by the tipping point model as by the form-driven model. This is because, as mentioned above, the tipping point model can effectively function as a form-driven model when the parsing penalty is so high that the recognition of plurals is always form-driven. In those cases, we prefer the form-driven model, because it is less complex than the tipping point model. Given these results, we conclude that, with our approach, it is possible to correctly infer the model that generated the RT distribution.

Some of the datasets that we analyse in the subsequent experiments contain fewer data points than the datasets tested in Study 1. It is therefore important to note that we established that the same conclusions are reached when Study 1 is carried out with smaller sets, consisting of 2,000 RTs (instead of 6,000 RTs), produced by 20 synthesised participants, responding to 100 synthesised words (instead of 300 synthesised words). That is, we found again that the model that generated the RTs makes the most accurate predictions of the RT distributions (Figure B.3). The exceptions are, again, the RTs generated by the form-driven model, which are predicted as accurately by the tipping point model as by the form-driven model.

One of the datasets that we analyse in Study 1 contains so many observations that it is infeasible for us to include all the data in the analysis, because the computation of the ELPD LOOs for this dataset exceeds the working memory capacity of 512 GB of our largest computing node. Following Magnusson et al. (2020), we compute the ELPD LOO on the basis of a random subsample of the data that consists of one tenth of all observations. To show that model comparisons are stable regardless of whether the ELPD LOO is based on the whole dataset or on a subset of 10%, we computed the ELPD LOO for the first dataset that was generated with the form-driven model both on the basis of all data points in that set and on the basis of ten different subsamples, each consisting of one tenth of the whole dataset. Figure B.4 shows that the difference in ELPD LOO between the three inferential models is approximately the same regardless of whether the ELPD LOO is based on all observations (first column) or on one of ten different subsets (other columns).

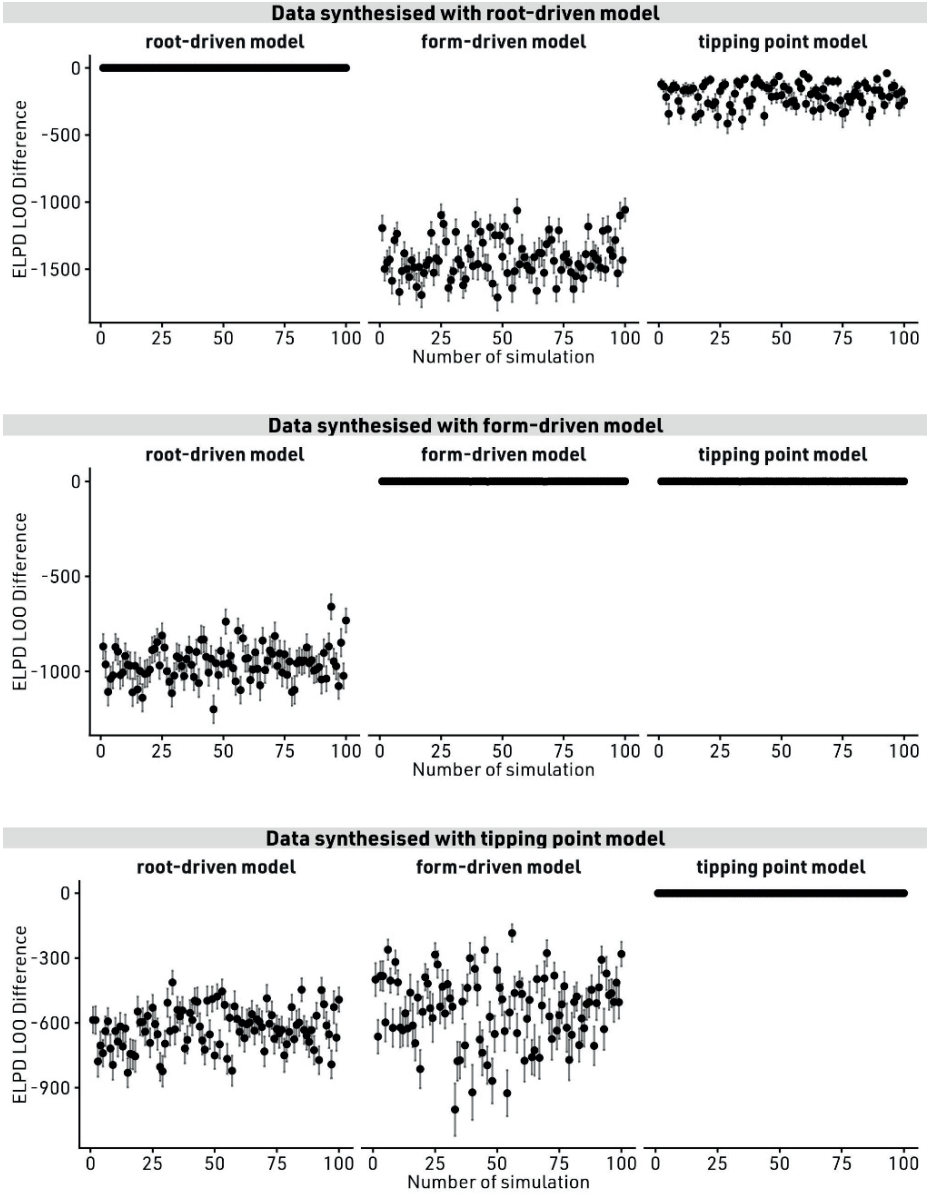


Figure B.2 Difference in ELPD LOO (y-axis) between the root-driven model (first column, the form-driven model (second column), and the tipping point model (third column) relative to the model with the highest ELPD LOO, for each synthesised distribution of 6,000 response times (x-axis), when the data was synthesised by the root-driven model (top row), the form-driven model (second row), and the tipping point model (bottom row). Error bars represent two times the standard error. The figure shows that the model that synthesised the data also fits the data the most accurately.

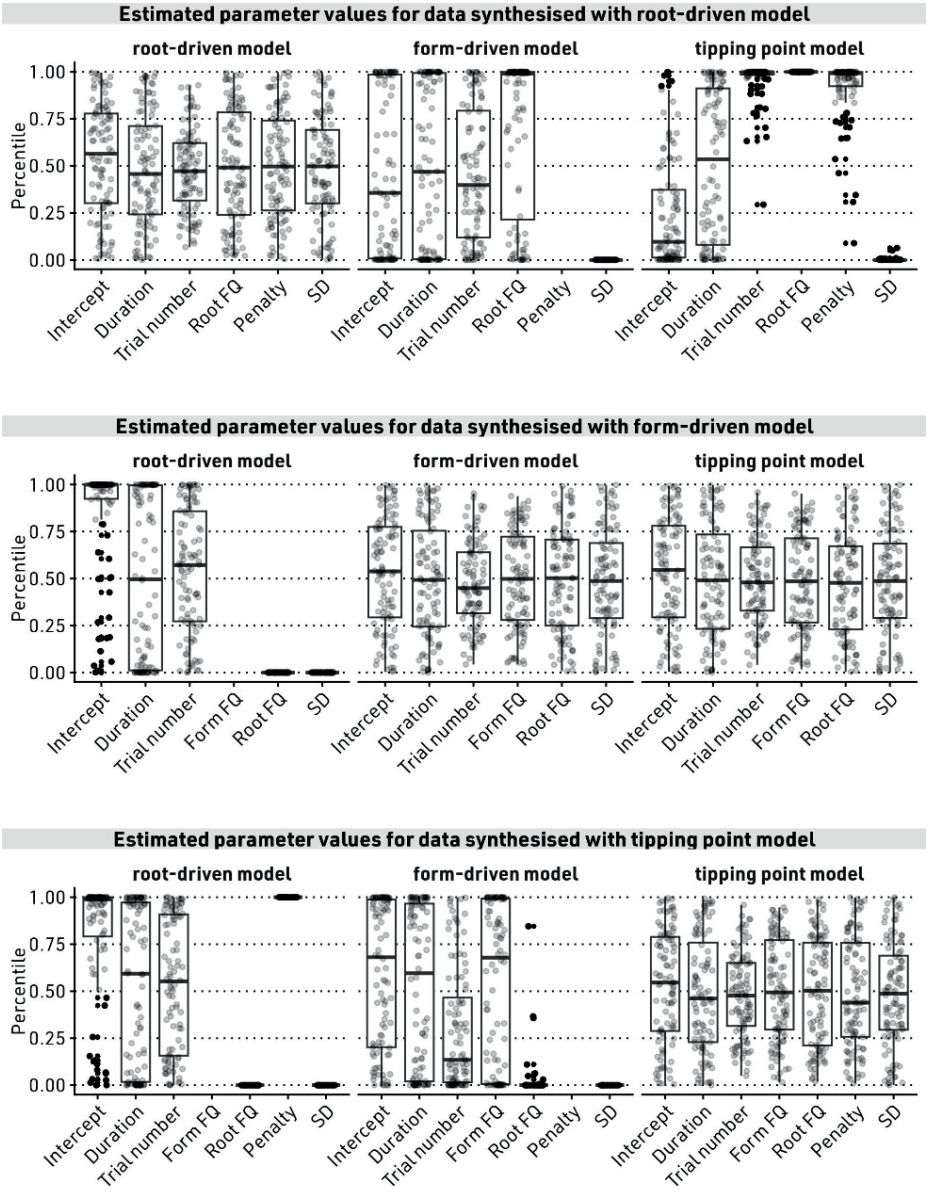


Figure B.3 Difference in ELPD LOO (y-axis) between the root-driven model (first column), the form-driven model (second column), and the tipping point model (third column) relative to the model with the highest ELPD LOO, for each synthesised distribution of 2,000 response times (x-axis), when the data was synthesised by the root-driven model (top row), the form-driven model (second row), and the tipping point model (bottom row). Error bars represent two times the standard error. The figure shows that the model that synthesised the data also fits the data the most accurately.

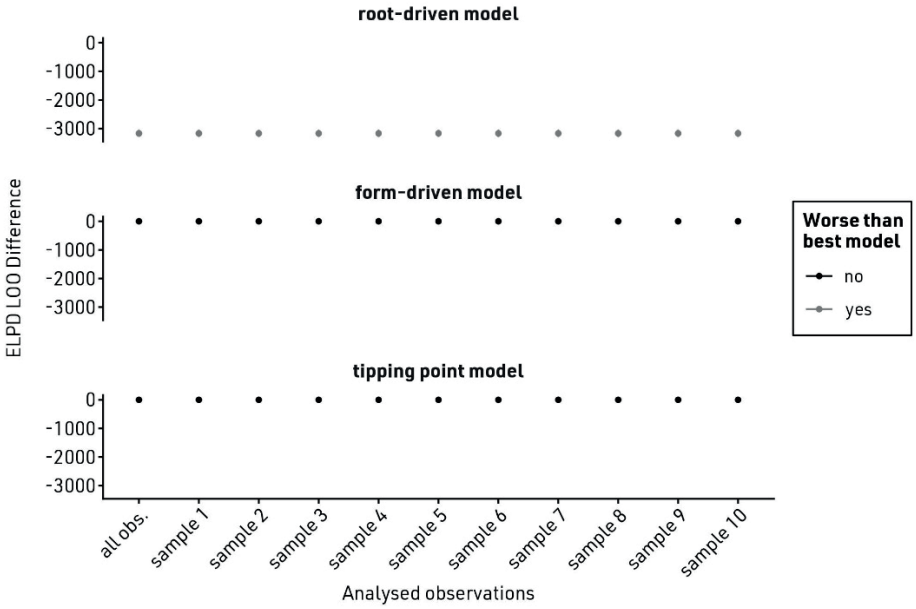


Figure B.4 Difference in ELPD LOO (y-axis) between the root-driven model (first row), the form-driven model (second row), and the tipping point model (third row) relative to the model with the highest ELPD LOO (which is for each mean parsing penalty the form-driven model and the tipping point model), for all observations of a given dataset (x-axis, very first column) and ten random subsamples (x-axis, other columns with labels sample 1, sample 2, ..., sample 10). The two horizontal dotted lines in each panel represent differences in ELPD LOO of zero and minus four. Error bars represent two times the standard error. The figure shows that the difference in ELPD LOO between the three models does not change when based on a subset of the data.

In conclusion, the simulation experiment shows that, although the RT distributions generated by the three models are highly correlated, it is possible to infer the generating model, in the controlled environment of generated RTs. Because we generated RT distributions based on an RT distribution of nouns in BALDEY (Ernestus & Cutler, 2015), we are confident that it is also possible to infer the processing model that best fits human lexical decision data. This validates our methodological approach.

Appendix C

For determining the parsing penalty prior for our models, we assumed that parsing should take at least 1 ms but no longer than 350 ms. Because all models were fitted with log-transformed predictors, we defined the parsing penalty prior on the log scale as well, that is, as $\Delta p_{log} = log(e^{Intercept} + \Delta p) - Intercept$, whereby the Intercept refers to the mean prior of the intercept, being 6.5173 (677 ms) for DLP (see Table 2.8) and 5.7090 (302 ms) for BALDEY (see Table 2.9). This implies that, on the log scale, we expect the mean of the parsing penalty to be between 0.0015 (1 ms) and 0.4168 (350 ms) for DLP and between 0.0033 (1 ms) and 0.7704 (350 ms) for BALDEY.

We tested sequences of means, of which some were slightly outside the range of plausible parsing penalties, because this enables us to investigate the model's estimations when provided with an implausible parsing penalty prior. Specifically, we explored normal distributions for the parsing penalty prior with means ranging from 0.001 to 0.801 in 0.02 increments while the corresponding standard deviations (SD) were defined by formula (2.4).

Note that the computed standard deviation tends to be a fourth of the mean for large values of the mean, with a lower bound of 1, which is attained for low values of the mean just above 0 (e.g., mean of 0.001 has a SD of $0.001 / 4 + 4 / (0.001 + 4) = 1$). This is necessary to ensure that priors smaller than 4 ms are not too informative so that updating the prior distribution based on the actual data would be impossible.

*

Appendix D

Table D.1 *Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 1 (written plural nouns ending in* -en*). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with FQ.*

	Root-driven model			Form-driven model			Tipping point model		
Estimate	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**
Intercept	6.51000	0.02170	1.00	6.52000	0.02140	1.01	6.53000	0.02060	1.01
Trial	0.00814	0.00087	1.00	0.00820	0.00087	1.00	0.00818	0.00087	1.00
Length	−0.00462	0.00176	1.00	−0.00407	0.00176	1.00	−0.00292	0.00177	1.00
Form FQ	–	–	–	−0.02410	0.00054	1.00	-0.02610	0.00060	1.00
Root FQ	−0.02640	0.00046	1.00	−0.02690	0.00041	1.00	−0.02910	0.00045	1.00
Penalty	0.06340 (0.0636)	0.00175	1.00	–	–	–	0.08300 (0.082)	0.00374	1.00

Table D.2 *Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 2 (spoken plural nouns ending in* -en*). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with FQ.*

	Root-driven model			Form-driven model			Tipping point model		
Estimate	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**
Intercept	5.85000	0.12200	1.01	5.88000	0.01310	1.01	5.90000	0.13300	1.01
Trial	0.01340	0.00253	1.00	0.01340	0.00261	1.00	0.01340	0.00256	1.00
Duration	0.18700	0.01730	1.01	0.18300	0.01780	1.01	0.18200	0.01780	1.00
Form FQ	–	–	–	-0.00524	0.00143	1.00	−0.00528	0.00145	1.00
Root FQ	−0.00632	0.00121	1.00	−0.00812	0.00111	1.00	−0.00813	0.00113	1.00
Penalty	0.03690 (0.0371)	0.00834	1.00	–	–	–	0.73400 (0.582)	0.04070	1.00

Table D.3 *Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 3 (written second/third person singular present tense verb forms ending in -t). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with FQ.*

	Root-driven model			Form-driven model			Tipping point model		
Estimate	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**
Intercept	6.42000	0.02300	1.00	6.43000	0.02200	1.01	6.42000	0.02230	1.01
Trial	0.00709	0.00145	1.00	0.00712	0.00146	1.00	0.00714	0.00147	1.00
Length	0.00613	0.00185	1.00	−0.00610	0.00188	1.00	0.00607	0.00186	1.00
Form FQ	–	–	–	−0.01110	0.00106	1.00	−0.01110	0.00106	1.00
Root FQ	−0.00781	0.00083	1.00	−0.00818	0.00074	1.00	−0.00819	0.00073	1.00
Penalty	0.01250 (0.0124)	0.00415	1.00	–	–	–	0.43900 (0.044)	0.26200	1.00

Table D.4 *Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 4 (spoken second/third person singular present tense verb forms ending in -t). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with FQ.*

	Root-driven model			Form-driven model			Tipping point model		
Estimate	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**
Intercept	5.60000	0.17200	1.01	5.50000	0.16500	1.00	5.49000	0.17000	1.01
Trial	0.01380	0.00282	1.00	0.01380	0.00274	1.00	0.01380	0.00271	1.00
Duration	0.22900	0.02500	1.01	0.24500	0.02440	1.00	0.24500	0.02490	1.00
Form FQ	–	–	–	−0.01230	0.00186	1.00	−0.01230	0.00200	1.00
Root FQ	−0.00593	0.00124	1.00	−0.00472	0.00107	1.00	−0.00469	0.00114	1.00
Penalty	0.00791 (0.0045)	0.00431	1.00	–	–	–	0.75100 (1.15)	0.41100	1.00

*

Table D.5 *Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 5 (written nouns ending in the derivational suffix -heid). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with FQ.*

	Root-driven model			Form-driven model			Tipping point model		
Estimate	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**
Intercept	6.44000	0.02110	1.01	6.45000	0.02110	1.00	6.45000	0.02070	1.01
Trial	0.00769	0.00152	1.00	0.00762	0.00150	1.00	0.00763	0.00151	1.00
Length	0.00568	0.00185	1.00	0.00509	0.00187	1.00	0.00507	0.00187	1.00
Form FQ	–	–	–	−0.00114	0.00144	1.00	−0.01140	0.00148	1.00
Root FQ	−0.01470	0.00085	1.00	−0.00163	0.00083	1.00	−0.01620	0.00083	1.00
Penalty	0.07870 (0.0797)	0.01470	1.00	–	–	–	0.50000 (0.526)	0.22800	1.00

Table D.6 *Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 6 (spoken nouns ending in the derivational suffix -heid). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with FQ.*

	Root-driven model			Form-driven model			Tipping point model		
Estimate	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**	**Mean**	**SD**	**Rhat**
Intercept	5.5400	0.16900	1.01	5.29000	0.15100	1.00	5.28000	0.14400	1.00
Trial	0.0123	0.00262	1.00	0.01220	0.00256	1.00	0.01240	0.00251	1.00
Duration	0.2350	0.02480	1.01	0.27600	0.02150	1.00	0.27700	0.02060	1.00
Form FQ	–	–	–	−0.01150	0.00178	1.00	−0.01160	0.00179	1.00
Root FQ	−0.00585	0.00116	1.00	-0.00620	0.00120	1.00	0.00618	0.00115	1.00
Penalty	0.02400 (0.237)	0.04410	1.00	–	–	–	0.72200 (0.913)	0.40700	1.00

*

Appendix E

Table E.1 *Summary of the baseline model plus Semantic Density and Semantic Form Overlap Family Size and interactions, fitted to log-transformed RTs in our subset of BALDEY.*

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	6.22343	0.03595	173.127	< .001
MorphStrprefixed	0.03174	0.04600	0.690	.490
MorphStrsuffixed	−0.02794	0.01152	−2.837	.015
MorphStrdouble-affixed	−0.15014	0.05292	−2.424	.005
B. smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(SemDensity:MorphStrsimplex)	1.855	1.979	75.543	< .001
s(SemDensity:MorphStrprefixed)	1.372	1.606	5.076	.006
s(SemDensity:MorphStrsuffixed)	1.978	1.999	286.310	< .001
s(SemDensity:MorphStrdouble-affixed)	1.000	1.000	0.006	.938
s(FamilySize:MorphStrsimplex)	1.000	1.000	20.716	< .001
s(FamilySize:MorphStrprefixed)	1.000	1.000	6.809	.009
s(FamilySize:MorphStrsuffixed)	8.116	8.784	6.005	< .001
s(FamilySize:MorphStrdouble-affixed)	1.000	1.000	12.308	< .001
s(Freq)	4.805	5.992	6.848	< .001
s(maRT)	4.869	6.084	302.558	< .001
s(Trial)	6.376	7.534	4.595	< .001
s(participant)	18.549	19.000	27.955	< .001
s(participant, Trial)	15.636	19.000	8.214	< .001

*

Appendix F

As explained in Chapter 5.4.2 and as illustrated in Figure 5.1, we augmented DIANA with a) morpheme representations in the lexicon, b) information on the relationships between morpheme and word representations, and c) a mechanism that updates morpheme probabilities based on word probabilities and, vice versa, that updates word probabilities based on morpheme probabilities, taking word–morpheme relationships into account. Here, we illustrate how we did so, on the basis of a case in which a 20 ms long word token of *three* is uttered, and the lexicon just contains three words (*three*, *threefold*, *fold*) and two morphemes (*three*, *fold*). Moreover, for simplicity, we assume that in the course of the word recognition process, only three pseudowords are activated (*thu*, *tri* and *sri*).

The word, morpheme, and pseudoword hypotheses form matrices (W, M, and N, respectively, see F.1):

$$\mathbf{W} = \begin{pmatrix} W_{1,1} \\ W_{2,1} \\ W_{3,1} \end{pmatrix} \begin{matrix} three \\ threefold \\ fold \end{matrix}, \quad \mathbf{M} = \begin{pmatrix} M_{1,1} \\ M_{2,1} \end{pmatrix} \begin{matrix} three \\ fold \end{matrix}, \quad \mathbf{N} = \begin{pmatrix} N_{1,1} \\ N_{2,1} \\ N_{3,1} \end{pmatrix} \begin{matrix} thu \\ tri \\ sri \end{matrix} \tag{F.1}$$

The morphological relationships between the word hypotheses and the morphemes are represented by matrix L, with w columns, corresponding to the words, and m rows, corresponding to the morphemes. Whenever a morpheme is part of a given word, the cell identified by this word and this morpheme has a value of 1, and otherwise 0 (see F.2).

$$\mathbf{L} = \begin{matrix} three & threefold & fold \\ \left(\begin{matrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{matrix} \right) & \begin{matrix} three \\ fold \end{matrix} \end{matrix} \tag{F.2}$$

Column normalisation ensures that the values in one column sum up to one (see F.3). This is necessary so that, through spreading activation, words consisting of multiple morphemes receive the same maximum amount of activation as words consisting of fewer morphemes.

$$\mathbf{L} \odot (\sum_{i=1}^{m} L_{i,w})^{-1} = \begin{matrix} three & threefold & fold \\ \left(\begin{matrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{matrix} \right) & \begin{matrix} three \\ fold \end{matrix} \end{matrix} \tag{F.3}$$

The word (WDIANA) and pseudoword (NDIANA) probabilities at t_{10} and t_{20} as provided by DIANA's activation component just on the basis of the audio signal are shown in Equations (F.4) and (F.5), respectively. See ten Bosch et al., (2022) for how these probabilities are computed.

$$\mathbf{WDIANA^{10}} = \begin{pmatrix} 0.2 \\ 0.1 \\ 0 \end{pmatrix} \begin{matrix} three \\ threefold \\ fold \end{matrix}, \quad \mathbf{NDIANA^{10}} = \begin{pmatrix} 0.3 \\ 0.2 \\ 0.2 \end{pmatrix} \begin{matrix} thu \\ tri \\ sri \end{matrix} \tag{F.4}$$

$$\mathbf{WDIANA^{20}} = \begin{pmatrix} 0.4 \\ 0.2 \\ 0 \end{pmatrix} \begin{matrix} three \\ threefold \\ fold \end{matrix}, \quad \mathbf{NDIANA^{20}} = \begin{pmatrix} 0.2 \\ 0.1 \\ 0.1 \end{pmatrix} \begin{matrix} thu \\ tri \\ sri \end{matrix} \tag{F.5}$$

Every time these word and pseudoword probabilities are updated on the basis of the acoustic signal, spreading activation takes place. That is, the bottom-up word probabilities given the acoustic signal DIANAt and the top-down word probabilities from spreading activation M^{t-1} are combined into the matrix W^t, by weighing the contribution of both types of probabilities as a function of the parameter *sensitivity s*, which lies in the interval between 0 and 1, (see F.6).

$$\mathbf{W^t} = s \times \mathbf{WDIANA^t} + (1 - s) \times \mathbf{M^{t-1}} \times \mathbf{L} \odot (\sum_{i=1}^{m} L_{i,w})^{-1} \tag{F.6}$$

To ensure that the summed pseudoword and word probabilities amount to 1, both probabilities are normalised. After the word probabilities have been updated, the morpheme probabilities at the current time stamp t (M^t) are also updated on the basis of 1) the morpheme probabilities from the previous time step M^{t-1}, 2) the word probabilities at the current time step W^t, and 3) the *decay* parameter *d*, which has a value between 0 and 1, (see F.7).

$$\mathbf{M^t} = d \times \mathbf{M^{t-1}} + (1 - d) \times \mathbf{W^t} \times \mathbf{L}^\intercal \tag{F.7}$$

Research Data Management

Personal data

Did you process data that has already been published? If yes, how did you ensure compliance with legislation on privacy?

Yes, I processed data that has already been published. I analysed data from the Dutch Biggest Auditory Lexical Decision Experiment Yet, the Dutch Lexicon Project, and the Corpus Gesproken Nederlands. The published data was anonymised or pseudo-anonymised by the right holders of these datasets, that is, participants are represented by numerical identifiers, because of which compliance with legislation on privacy is ensured.

Did you process personal data? If yes, how did you ensure compliance with legislation on privacy?

Yes, during my research, I processed personal data in form of the above-mentioned datasets. Because these data are (pseudo-)anonymised, compliance with legislation on privacy is ensured.

In order to protect the privacy of your participants, did you anonymise or pseudo-anonymise the data?

I did not collect data and thus I did not have to deal with personal information of participants.

Did you need approval from an ethics committee for your project?

No, for my research I did not need approval from an ethics committee.

Did your research require an informed consent procedure?

No, my research did not require an informed consent procedure.

Storing and sharing during research

Did you make use of safe storage during your research, including back-up facilities?

Yes, I followed the policy of my institute and stored the data on the Radboud Data Repository (RDR; http://data.ru.nl), which guarantees access to stored data for at least 10 years. In addition, I stored scripts for data processing on the university's network drive Ponyland and a private GitHub repository. I created back-ups of the scripts on a daily basis, using the GitHub repository.

When off-campus, I securely accessed the data through a VPN connection. I made sure that the structure of the data on the RDR, Ponyland, and GitHub meet the minimum requirements as described in my institute's research data management (RDM) protocol.

With whom did you share your data during research?

During my research, I did not share data with other researchers.

How did you deal with security issues that arose during your research?

During my research, the data has been stored on the university's network drive Ponyland. This storage location meets legal and ethical requirements. Safe and secure storage was guaranteed by the IT security and safety protocols.

I organised my project's folder according to the following format:

I organised the structure of my research data folder according to the minimum requirements as described in my institute's RDM protocol.

*

Long term archiving and reuse

In the context of scientific integrity, where did you archive your data (including raw data, metadata and documentation) for at least 10 years?

I followed my institute's policy and ensured to archive the research data associated with my publication (including metadata and documentation) in a Data Sharing Collection (DSC) in the RDR for a minimum of 10 years. I did not add the raw data to the DSC, because the raw data can be downloaded from the internet. In addition, the raw data is stored on the university's network drive Ponyland.

In the context of data reuse, did you make your research data publicly available?

Yes, I followed my institute's policy and made my data public at publication via the RDR. The data include metadata, documentation, and the pre-processed datasets that I analysed in my studies. The data will be published using one of the available open access licences.

How did you ensure that your research data will be stored in a FAIR manner?

I ensured that my research data will be stored in a FAIR manner by using the RDR for data publication and archiving. My research data is findable through a DSC in the RDR, which is indexed by search engines. My data includes rich metadata and a persistent identifier (DOI). Because the RDR uses open internet protocol and clear authorization procedures, my research data is accessible. My research data is interoperable, because I used standards for metadata and standard preferred data formats. Because of a clear license, rich metadata, and documentation, my research data is reusable through the RDR.

Which datasets did you publish through the Radboud Data Repository?

I created four DSCs on the RDR:

1) The DSC called "A tipping point in word recognition? Investigating the relationship between root and form frequency across visual and auditory modalities." is with following DOI: https://doi.org/10.34973/jm3a-vj10
2) The DSC called "The Family Size Effect in Visual and Auditory Word Recognition" is published with the following DOI: https://doi.org/10.34973/yc4q-r262
3) The DSC called "Can Discriminative Lexicon Theory account for the family size effect in auditory word recognition?" is published with the following DOI: https://doi.org/10.34973/x6v3-yj45
4) The DSC called "How well does a spreading activation mechanism account for the auditory family size effect?" is published with the following DOI: https://doi.org/10.34973/71n3-vk61

Each DSC contains an abstract of the study associated with the data. Above that, each DSC contains a read-me file describing the data, the pre-processing steps, and the final analysis of the pre-processed data, as well as the same folder structure: scripts for pressing the raw data can be found in the folder "ScriptsProcessing"; the dataset that is analysed in the study is stored in the folder "DataProcessed"; scripts for analysing this dataset are provided in the folder "ScriptsAnalysis".

*

English Summary

Understanding spoken language is a central part of human communication. In everyday situations, people recognise spoken words quickly and with little apparent effort, whether in casual conversation or when listening to more abstract content such as the news. This apparent ease, however, hides a complex cognitive problem. Spoken words are conveyed through acoustic signals that change continuously over time, while the words are usually perceived as relatively stable and discrete units. Research on spoken word recognition aims to explain how listeners manage to map these unfolding signals onto words. The thesis summarised here addresses this question by examining the role of morphology in word recognition, that is, the role of the internal structure of words and their relations to other words in the mental lexicon.

Many words are morphologically complex (e.g., *viewer*) and consist of smaller meaningful units, such as roots (e.g., *view*) and affixes (e.g., *-er*). In research on reading, there is strong evidence that such morphological structure influences how written words are processed. For spoken words, however, the role of morphology is much less clear. One reason is that spoken and written word recognition differ in important ways. Written words are typically visible all at once and can often be processed as complete forms, whereas spoken words are recognised incrementally as the acoustic signal unfolds. For this reason, findings from reading cannot simply be applied to listening. This thesis aims to clarify whether and how morphological structure affects spoken word recognition and which theoretical mechanisms can best explain the observed effects.

This thesis has three main aims. First, it investigates whether the morphological structure of spoken words influences how quickly listeners recognise them. Second, it asks which cognitive mechanisms might underlie such morphological effects. Third, it explores what can be learned by implementing and comparing explicit computational models that represent different theoretical views of word recognition. In this thesis, Morphology is therefore not treated only as a linguistic description of word structure, but as a factor that may shape observable behaviour and the internal dynamics of formal models implementing different word recognition theories.

To address these questions, the thesis uses a combination of lexical decision experiments and computational modelling. In a lexical decision experiment, participants decide as quickly as possible whether a stimulus is a real word or not. How quick a decision is made, is measured as response time in ms. This paradigm is widely used in research on word recognition and has the advantage that the recognition of words can be investigated in isolation, that is, without sentence context. In addition, large lexical decision datasets are available for both spoken and written words. These large datasets make it possible to conduct robust statistical analyses and to compare results across modalities.

The empirical core of the thesis consists of analyses of large Dutch auditory and visual lexical decision datasets. Response times are analysed using statistical models that test whether specific predictors systematically explain differences in how quickly words are recognised. These predictors are either properties of the words themselves, such as different types of frequency (i.e., how often a word occurs in a corpus), or outputs of computational models that embody entire theories of word recognition. This approach does not only make it possible to identify morphological effects, but also to compare competing theoretical explanations.

One major focus is the relative importance of roots and whole-word forms in recognising morphologically complex words. Such words have a surface form frequency, reflecting how often the full word (e.g., *viewer*) occurs, and a root frequency, reflecting how often the root (e.g., *view*) appears across related words (e.g., *preview*, *views*). Different theories make different predictions about which type of frequency should matter most. Some assume that recognition is mainly driven by roots, others that it is driven by whole-word forms, and still others that both contribute. By testing statistical models corresponding to these positions, the thesis shows that neither a purely root-driven nor a purely form-driven account can explain all observed patterns. Instead, word recognition appears to lie on a continuum: surface form frequency usually plays the dominant role, but root frequency can contribute under certain conditions.

Importantly, the results indicate that the balance between root-driven and form-driven processing is gradual rather than categorical. For some types of words, especially in written word recognition and for certain suffixes, roots exert a stronger influence. For others, particularly in spoken word recognition, whole-word forms dominate. The relative contribution of roots and forms cannot be predicted by frequency ratios alone. Other factors, such as whether words are heard or read and the properties of specific affixes, also shape how morphological information is used. These findings show that morphology matters in spoken word recognition, but in a nuanced way.

A second major theme of the thesis is the morphological family size effect. Morphological family size refers to the number of words that share a common root, such as *view*, *viewer*, and *preview*. In reading research, words with larger families are typically recognised faster, often explained by shared semantic or morphological representations. For spoken words, previous findings have been inconsistent. By analysing very large auditory and visual lexical decision datasets, the thesis shows clearly that spoken words also show a facilitative family size effect: words with more morphologically related neighbours are recognised more quickly.

The effect is not the same for all word types. Spoken simplex words (e.g., *view*) and suffixed words (e.g., *viewer*) show a clear family size effect, whereas prefixed words (e.g., *review*) do not. This difference is explained by the time course of spoken word recognition. In prefixed words, the root becomes available relatively late in the signal, which limits the influence of related words during early processing. In written word recognition, by contrast, prefixes and roots are often visible at the same time, making family size effects more likely to emerge.

The thesis also examines what properties of family members drive the family size effect. Different measures are compared, varying in whether they emphasise semantic similarity, form overlap, or both. The analyses show that the effect is strongest when family members are weighted by both how similar they are in meaning and how similar they are in form. Semantic similarity plays an important role, but form overlap also contributes in a graded way rather than as a simple yes–or–no factor. By using continuous measures of phonological and semantic similarity, the thesis provides a more fine-grained account of how related words influence recognition.

Beyond behavioural analyses, the thesis evaluates how well different theoretical approaches can explain the observed effects. Two broad classes of theories are considered. *Distributional–connectionist* theories view word recognition as the result of learned mappings between form and meaning, without assuming explicit representations of morphemes. *Localist* theories, by contrast, assume discrete representations for words and morphemes and often explain family size effects through spreading activation between related units. The thesis implements computational models representing each approach and tests how well they predict auditory lexical decision response times.

On the distributional–connectionist side, an auditory version of the Discriminative Lexicon Model is applied. This model predicts response times based on how well acoustic input can be mapped onto semantic representations. The model accounts for part of the auditory family size effect, showing that some aspects of the effect can arise from general learning mechanisms that exploit shared form–meaning structure, even without explicit morpheme representations. However, it does not fully explain the effect.

On the localist side, the spoken word recognition model DIANA is used. DIANA generates and updates hypotheses about possible words as the acoustic signal unfolds. Measures derived from these hypotheses explain part of the family size effect, even though the model does not explicitly encode morphological relationships. This is attributed to the fact that words with many family members often have many similar-sounding neighbours, which affects the model's hypothesis space. Adding an explicit spreading activation mechanism to DIANA, intended to model morphological relationships, does not improve its performance.

Overall, the two modelling approaches perform similarly once family members are weighted by form and semantic similarity. The thesis therefore concludes that current evidence does not clearly favour one mechanism over the other.

Finally, the thesis reflects on the role of computational modelling in research on spoken word recognition. Implementing theories as explicit models forces assumptions to be stated clearly and reveals that concepts such as frequency, similarity, or morphological structure need careful definition. At the same time, models are constrained by the data and tasks they are applied to. Lexical decision experiments offer valuable insights but also have clear limitations. The thesis therefore calls for future work that combines modelling with other

methods, such as neurophysiological measures, to better capture the time course of spoken word recognition.

In summary, the thesis shows that morphological structure influences the recognition of spoken words, but in a graded and word-dependent manner. Both root-based and whole-word information contribute, and spoken words show a clear morphological family size effect under specific conditions. By combining large-scale behavioural data with explicit computational models, the work demonstrates that morphological effects do not necessarily require explicit morpheme representations, but may also arise from more general properties of the lexical system. Together, these findings advance our understanding of how listeners recognise morphologically complex spoken words and highlight the value of integrating empirical data with formal modelling.

*

Nederlandse Samenvatting

Het begrijpen van gesproken taal is een centraal onderdeel van menselijke communicatie. In alledaagse situaties herkennen mensen gesproken woorden snel en ogenschijnlijk moeiteloos, zowel in informele gesprekken als bij het luisteren naar meer abstracte inhoud, zoals het nieuws. Deze schijnbare eenvoud verhult echter een complex cognitief probleem. Gesproken woorden worden overgebracht via akoestische signalen die continu in de tijd veranderen, terwijl de woorden die luisteraars waarnemen relatief stabiele en discrete eenheden zijn. Onderzoek naar gesproken woordherkenning probeert te verklaren hoe luisteraars erin slagen deze zich ontvouwende signalen aan woorden te koppelen. Het hier samengevatte proefschrift behandelt deze vraag door de rol van morfologie bij woordherkenning te onderzoeken, dat wil zeggen de interne structuur van woorden en hun relaties tot andere woorden in het mentale lexicon.

Veel woorden zijn morfologisch complex (bijv. *kijker*) en bestaan uit kleinere betekenisvolle eenheden, zoals stammen (bijv. *kijk*) en affixen (bijv. *-er*). In onderzoek naar lezen is er sterk bewijs dat dergelijke morfologische structuur invloed heeft op de verwerking van geschreven woorden. Voor gesproken woorden is de rol van morfologie echter veel minder duidelijk. Een belangrijke reden hiervoor is dat gesproken en geschreven woordherkenning op essentiële punten van elkaar verschillen. Geschreven woorden zijn doorgaans in één keer zichtbaar en kunnen vaak als volledige vormen worden verwerkt, terwijl gesproken woorden incrementeel worden herkend naarmate het akoestische signaal zich ontvouwt. Om deze reden kunnen bevindingen uit leesonderzoek niet eenvoudig worden toegepast op luisteren. Het proefschrift heeft tot doel te verduidelijken of en hoe morfologische structuur de herkenning van gesproken woorden beïnvloedt en welke theoretische mechanismen de waargenomen effecten het best kunnen verklaren.

Het proefschrift heeft drie hoofddoelen. Ten eerste onderzoekt het of de morfologische structuur van gesproken woorden invloed heeft op hoe snel luisteraars deze herkennen. Ten tweede wordt gevraagd welke cognitieve mechanismen aan dergelijke morfologische effecten ten grondslag zouden kunnen liggen. Ten derde wordt verkend wat kan worden geleerd door expliciete computationele modellen te implementeren en te vergelijken

die verschillende theoretische visies op woordherkenning representeren. Morfologie wordt daarmee niet alleen beschouwd als een linguïstische beschrijving van woordstructuur, maar ook als een factor die observeerbaar gedrag en de interne dynamiek van formele modellen die verschillende woordherkenningstheorieën implementeren, kan vormgeven.

Om deze vragen te beantwoorden, maakt het proefschrift gebruik van een combinatie van lexicale beslissingsexperimenten en computationele modellering. In een lexicaal beslissingsexperiment beslissen deelnemers zo snel mogelijk of een stimulus een echt woord is of niet. Hoe snel een beslissing wordt genomen, wordt gemeten als reactietijd in milliseconden. Dit paradigma wordt veel gebruikt in onderzoek naar woordherkenning en heeft als voordeel dat de herkenning van woorden in isolatie kan worden onderzocht, zonder zinscontext. Daarnaast zijn er grote lexicale beslissingsdatasets beschikbaar voor zowel gesproken als geschreven woorden. Deze grote datasets maken het mogelijk robuuste statistische analyses uit te voeren en resultaten over modaliteiten heen te vergelijken.

De empirische kern van het proefschrift bestaat uit analyses van grote Nederlandse auditieve en visuele lexicale beslissingsdatasets. Reactietijden worden geanalyseerd met behulp van statistische modellen die toetsen of specifieke voorspellers systematisch verschillen in herkenningssnelheid van woorden verklaren. Deze voorspellers zijn ofwel eigenschappen van de woorden zelf, zoals verschillende typen frequentie (dat wil zeggen hoe vaak een woord in een corpus voorkomt), ofwel uitkomsten van computationele modellen die volledige theorieën over woordherkenning belichamen. Deze aanpak maakt het niet alleen mogelijk morfologische effecten te identificeren, maar ook concurrerende theoretische verklaringen te vergelijken.

Een belangrijk aandachtspunt is het relatieve belang van stammen en volledige woordvormen bij het herkennen van morfologisch complexe woorden. Dergelijke woorden hebben een oppervlaktevormfrequentie, die weergeeft hoe vaak het volledige woord (bijv. *kijker*) voorkomt, en een stamfrequentie, die weergeeft hoe vaak de stam (bijv. *kijk*) voorkomt in verwante woorden (bijv. *terugkijken*, *kijkavond*). Verschillende theorieën doen verschillende

voorspellingen over welk type frequentie het belangrijkst zou moeten zijn. Sommige gaan ervan uit dat herkenning vooral door stammen wordt gestuurd, andere dat deze door volledige woordvormen wordt gedreven, en weer andere dat beide bijdragen. Door statistische modellen te schatten die overeenkomen met deze posities, laat het proefschrift zien dat noch een puur stamgedreven noch een puur vormgedreven benadering alle waargenomen patronen kan verklaren. In plaats daarvan lijkt woordherkenning op een continuüm te liggen: oppervlaktevormfrequentie speelt doorgaans de dominante rol, maar stamfrequentie kan onder bepaalde omstandigheden bijdragen.

Belangrijk is dat de resultaten aangeven dat de balans tussen stamgedreven en vormgedreven verwerking geleidelijk is en niet categorisch. Voor sommige typen woorden, vooral in geschreven woordherkenning en bij bepaalde suffixen, oefenen stammen een sterkere invloed uit. Voor andere, met name in gesproken woordherkenning, domineren volledige woordvormen. De relatieve bijdrage van stammen en vormen kan niet uitsluitend worden voorspeld op basis van frequentieverhoudingen. Andere factoren, zoals of woorden worden gehoord of gelezen en de eigenschappen van specifieke affixen, beïnvloeden eveneens hoe morfologische informatie wordt gebruikt. Deze bevindingen laten zien dat morfologie een rol speelt in gesproken woordherkenning, maar op een genuanceerde manier.

Een tweede belangrijk thema van het proefschrift is het effect van morfologische familiegrootte. Morfologische familiegrootte verwijst naar het aantal woorden dat een gemeenschappelijke stam deelt, zoals *kijk*, *kijker* en *terugkijken*. In leesonderzoek worden woorden met grotere families doorgaans sneller herkend, wat vaak wordt verklaard door gedeelde semantische of morfologische representaties. Voor gesproken woorden waren eerdere bevindingen inconsistent. Door zeer grote auditieve en visuele lexicale beslissingsdatasets te analyseren, laat het proefschrift duidelijk zien dat ook gesproken woorden een faciliterend familiegrootte-effect vertonen: woorden met meer morfologisch verwante woorden worden sneller herkend.

Het effect is niet voor alle woordtypen hetzelfde. Gesproken morfologisch eenvoudige woorden (bijv. *kijk*) en gesuffigeerde woorden (bijv. *kijker*) vertonen een duidelijk familiegrootte-effect, terwijl geprefixeerde woorden (bijv. *terugkijken*) dit niet doen. Dit verschil wordt verklaard door het tijdsverloop van gesproken woordherkenning. In geprefixeerde woorden wordt de stam relatief laat in het signaal beschikbaar, wat de invloed van verwante

woorden tijdens vroege verwerking beperkt. In geschreven woordherkenning daarentegen zijn prefixen en stammen vaak tegelijkertijd zichtbaar, waardoor familiegrootte-effecten eerder kunnen optreden.

Het proefschrift onderzoekt ook welke eigenschappen van familieleden het familiegrootte-effect aandrijven. Verschillende maten worden vergeleken, die variëren in de mate waarin zij semantische gelijkenis, vormoverlap of beide benadrukken. De analyses laten zien dat het effect het sterkst is wanneer familieleden worden gewogen op zowel betekenisgelijkenis als vormgelijkenis. Semantische gelijkenis speelt een belangrijke rol, maar vormoverlap draagt eveneens bij op een geleidelijke manier, in plaats van als een eenvoudige ja-nee-factor. Door gebruik te maken van continue maten voor fonologische en semantische gelijkenis biedt het proefschrift een fijnmaziger verklaring van hoe verwante woorden herkenning beïnvloeden.

Naast gedragsanalyses evalueert het proefschrift hoe goed verschillende theoretische benaderingen de waargenomen effecten kunnen verklaren. Twee brede klassen van theorieën worden beschouwd. *Distributioneel-connectionistische* theorieën beschouwen woordherkenning als het resultaat van aangeleerde koppelingen tussen vorm en betekenis, zonder expliciete representaties van morfemen te veronderstellen. *Localistische* theorieën daarentegen gaan uit van discrete representaties voor woorden en morfemen en verklaren familiegrootte-effecten vaak via spreidingsactivatie tussen verwante eenheden. Het proefschrift implementeert computationele modellen die elke benadering representeren en test hoe goed zij auditieve lexicale beslissingsreactietijden voorspellen.

Aan de distributioneel-connectionistische kant wordt een auditieve versie van het Discriminative Lexicon Model toegepast. Dit model voorspelt reactietijden op basis van hoe goed akoestische input kan worden gekoppeld aan semantische representaties. Het model verklaart een deel van het auditieve familiegrootte-effect en laat zien dat sommige aspecten van het effect kunnen voortkomen uit algemene leermechanismen die gebruikmaken van gedeelde vorm-betekenisstructuur, zelfs zonder expliciete morfeemrepresentaties. Het verklaart het effect echter niet volledig.

Aan de localistische kant wordt het model voor gesproken woordherkenning DIANA gebruikt. DIANA genereert en actualiseert hypothesen over mogelijke woorden naarmate het akoestische signaal zich ontvouwt. Maten die uit deze

hypothesen worden afgeleid verklaren een deel van het familiegrootte-effect, ook al codeert het model geen expliciete morfologische relaties. Dit wordt toegeschreven aan het feit dat woorden met veel familieleden vaak veel vergelijkbaar klinkende buren hebben, wat de hypotheseruimte van het model beïnvloedt. Het toevoegen van een expliciet spreidingsactivatiemechanisme aan DIANA, bedoeld om morfologische relaties te modelleren, verbetert de prestaties niet. Over het geheel genomen presteren de twee modelleringsbenaderingen vergelijkbaar zodra familieleden worden gewogen naar vorm- en semantische gelijkenis. Het proefschrift concludeert daarom dat het huidige bewijs niet duidelijk de ene mechanisme boven de andere verkiest.

Tot slot reflecteert het proefschrift op de rol van computationele modellering in onderzoek naar gesproken woordherkenning. Het implementeren van theorieën als expliciete modellen dwingt aannames duidelijk te formuleren en maakt zichtbaar waar concepten als frequentie, gelijkenis of morfologische structuur zorgvuldig moeten worden gedefinieerd. Tegelijkertijd worden modellen begrensd door de data en taken waarop zij worden toegepast. Lexicale beslissingsexperimenten bieden waardevolle inzichten, maar hebben ook duidelijke beperkingen. Het proefschrift pleit daarom voor toekomstig onderzoek dat modellering combineert met andere methoden, zoals neurofysiologische metingen, om het tijdsverloop van gesproken woordherkenning beter vast te leggen.

Samenvattend laat het proefschrift zien dat morfologische structuur de herkenning van gesproken woorden beïnvloedt, maar op een geleidelijke en woordafhankelijke manier. Zowel stamgebaseerde als volledige-woord-informatie dragen bij, en gesproken woorden vertonen onder specifieke omstandigheden een duidelijk morfologisch familiegrootte-effect. Door grootschalige gedragsdata te combineren met expliciete computationele modellen toont het werk aan dat morfologische effecten niet noodzakelijk expliciete morfeemrepresentaties vereisen, maar ook kunnen voortkomen uit meer algemene eigenschappen van het lexicale systeem. Samen verdiepen deze bevindingen ons begrip van hoe luisteraars morfologisch complexe gesproken woorden herkennen en onderstrepen zij de waarde van het integreren van empirische data met formele modellering.

*

Deutsche Zusammenfassung

Das Verstehen gesprochener Sprache ist ein zentraler Bestandteil menschlicher Kommunikation. In Alltagssituationen erkennen Menschen gesprochene Wörter schnell und mit geringem Aufwand, sei es in ungezwungenen Gesprächen oder beim Hören abstrakterer Inhalte wie der Nachrichten. Diese scheinbare Leichtigkeit verbirgt jedoch ein komplexes kognitives Problem. Gesprochene Wörter werden über akustische Signale vermittelt, die sich kontinuierlich über die Zeit verändern, während die Wörter, die Hörerinnen und Hörer wahrnehmen, relativ stabile und diskrete Einheiten sind. Die Forschung zur gesprochenen Worterkennung zielt darauf ab zu erklären, wie es Hörerinnen und Hörern gelingt, diese sich entfaltenden Signale auf Wörter abzubilden. Die hier zusammengefasste Dissertation geht dieser Frage nach, indem sie die Rolle der Morphologie bei der Worterkennung untersucht, also der internen Struktur von Wörtern und ihrer Beziehungen zu anderen Wörtern im mentalen Lexikon.

Viele Wörter sind morphologisch komplex (z. B. *kindlich*) und bestehen aus kleineren bedeutungstragenden Einheiten, wie Stämmen (z. B. *kind*) und Affixen (z. B. *-lich*). In der Leseforschung gibt es starke Hinweise darauf, dass eine solche morphologische Struktur die Verarbeitung geschriebener Wörter beeinflusst. Für gesprochene Wörter ist die Rolle der Morphologie jedoch deutlich weniger klar. Ein Grund dafür ist, dass sich gesprochene und geschriebene Worterkennung in wichtigen Aspekten unterscheiden. Geschriebene Wörter sind in der Regel auf einen Blick sichtbar und können oft als vollständige Formen verarbeitet werden, während gesprochene Wörter inkrementell erkannt werden, während sich das akustische Signal entfaltet. Aus diesem Grund lassen sich Befunde aus der Leseforschung nicht einfach auf das Hören übertragen. Ziel der Dissertation ist es, zu klären, ob und wie morphologische Struktur die gesprochene Worterkennung beeinflusst und welche theoretischen Mechanismen die beobachteten Effekte am besten erklären können.

Die Dissertation verfolgt drei Hauptziele. Erstens wird untersucht, ob die morphologische Struktur gesprochener Wörter beeinflusst, wie schnell Hörerinnen und Hörer sie erkennen. Zweitens wird gefragt, welche kognitiven Mechanismen solchen morphologischen Effekten zugrunde liegen könnten.

Drittens wird exploriert, was sich durch die Implementierung und den Vergleich expliziter computergestützter Modelle lernen lässt, die unterschiedliche theoretische Sichtweisen der Worterkennung repräsentieren. Morphologie wird damit nicht nur als linguistische Beschreibung der Wortstruktur betrachtet, sondern als ein Faktor, der beobachtbares Verhalten und die interne Dynamik formaler Modelle unterschiedlicher Worterkennungstheorien mitgestalten kann.

Zur Beantwortung dieser Fragen kombiniert die Dissertation lexikalische Entscheidungsexperimente mit computergestützter Modellierung. In einem lexikalischen Entscheidungsexperiment entscheiden Versuchspersonen so schnell wie möglich, ob ein Stimulus ein echtes Wort ist oder nicht. Die Geschwindigkeit der Entscheidung wird als Reaktionszeit in Millisekunden gemessen. Dieses Paradigma ist in der Worterkennungsforschung weit verbreitet und hat den Vorteil, dass die Erkennung von Wörtern isoliert, also ohne Satzkontext, untersucht werden kann. Zudem stehen große Datensätze zu lexikalischen Entscheidungen sowohl für gesprochene als auch für geschriebene Wörter zur Verfügung. Diese umfangreichen Datensätze ermöglichen robuste statistische Analysen und Vergleiche zwischen Modalitäten.

Der empirische Kern der Dissertation besteht aus Analysen großer niederländischer auditiver und visueller lexikalischer Entscheidungsdatensätze. Die Reaktionszeiten werden mithilfe statistischer Modelle analysiert, die prüfen, ob bestimmte Prädiktoren systematisch Unterschiede in der Geschwindigkeit der Worterkennung erklären. Diese Prädiktoren sind entweder Eigenschaften der Wörter selbst, wie verschiedene Arten von Frequenz (also wie häufig ein Wort in einem Korpus vorkommt), oder Ausgaben computergestützter Modelle, die vollständige Worterkennungstheorien verkörpern. Dieser Ansatz ermöglicht nicht nur die Identifikation von morphologischen Effekten, sondern auch den Vergleich konkurrierender theoretischer Erklärungen.

Ein zentrales Augenmerk liegt auf der relativen Bedeutung von Stämmen und Ganzwortformen bei der Erkennung morphologisch komplexer Wörter. Solche Wörter weisen eine Oberflächenformfrequenz auf, die angibt,

wie häufig das vollständige Wort (z. B. *kindlich*) vorkommt, sowie eine Stammfrequenz, die angibt, wie häufig der Stamm (z. B. *kind*) in verwandten Wörtern (z. B. *Kleinkind*, *kindlich*, *kindgerecht*) erscheint. Unterschiedliche Theorien machen unterschiedliche Vorhersagen darüber, welche Art von Frequenz am wichtigsten sein sollte. Einige gehen davon aus, dass die Erkennung hauptsächlich stammbasiert ist, andere nehmen an, dass sie von Ganzwortformen gesteuert wird, und wieder andere, dass beide beitragen. Durch die Anpassung statistischer Modelle, die diesen Positionen entsprechen, zeigt die Dissertation, dass weder ein rein stammbasiertes noch ein rein formbasiertes Modell alle beobachteten Muster erklären kann. Stattdessen scheint die Worterkennung auf einem Kontinuum zu liegen: Die Oberflächenformfrequenz spielt in der Regel die dominierende Rolle, doch die Stammfrequenz kann unter bestimmten Bedingungen ebenfalls beitragen.

Wichtig ist, dass die Ergebnisse darauf hindeuten, dass das Gleichgewicht zwischen stammbasierter und formbasierter Verarbeitung graduell und nicht kategorial ist. Für bestimmte Worttypen, insbesondere in der geschriebenen Worterkennung und bei bestimmten Suffixen, üben Stämme einen stärkeren Einfluss aus. Für andere, insbesondere in der gesprochenen Worterkennung, dominieren Ganzwortformen. Der relative Beitrag von Stämmen und Formen lässt sich nicht allein anhand von Frequenzverhältnissen vorhersagen. Weitere Faktoren, wie ob Wörter gehört oder gelesen werden und die Eigenschaften spezifischer Affixe, beeinflussen ebenfalls, wie morphologische Information genutzt wird. Diese Befunde zeigen, dass Morphologie in der gesprochenen Worterkennung eine Rolle spielt, jedoch auf differenzierte Weise.

Ein zweites zentrales Thema der Dissertation ist der Effekt der morphologischen Familiengröße. Die morphologische Familiengröße bezeichnet die Anzahl der Wörter, die einen gemeinsamen Stamm teilen, wie etwa *Kleinkind*, *kindlich* und *kindgerecht*. In der Leseforschung werden Wörter mit größeren Familien in der Regel schneller erkannt, was häufig durch geteilte semantische oder morphologische Repräsentationen erklärt wird. Für gesprochene Wörter waren frühere Befunde uneinheitlich. Durch die Analyse sehr großer auditiver und visueller lexikalischer Entscheidungsdatensätze zeigt die Dissertation jedoch eindeutig, dass auch gesprochene Wörter einen erleichternden Familiengrößeneffekt aufweisen: Wörter mit mehr morphologisch verwandten Nachbarn werden schneller erkannt.

Der Effekt ist jedoch nicht für alle Worttypen gleich. Gesprochene einfache (nicht abgeleitete) Wörter (z. B. *Kind*) und suffigierte Wörter (z. B. *kindlich*) zeigen einen klaren Familiengrößeneffekt, während präfigierte Wörter (z. B. *Kleinkind*) diesen nicht zeigen. Dieser Unterschied wird durch den zeitlichen Verlauf der gesprochenen Worterkennung erklärt. Bei präfigierten Wörtern wird der Stamm relativ spät im Signal verfügbar, was den Einfluss verwandter Wörter während der frühen Verarbeitung begrenzt. In der geschriebenen Worterkennung hingegen sind Präfixe und Stämme häufig gleichzeitig sichtbar, wodurch Familiengrößeneffekte eher auftreten können.

Die Dissertation untersucht zudem, welche Eigenschaften der Familienmitglieder den Familiengrößeneffekt antreiben. Es werden verschiedene Maße verglichen, die sich darin unterscheiden, ob sie semantische Ähnlichkeit, Formüberlappung oder beides betonen. Die Analysen zeigen, dass der Effekt am stärksten ist, wenn Familienmitglieder sowohl nach Bedeutungsähnlichkeit als auch nach Formähnlichkeit gewichtet werden. Semantische Ähnlichkeit spielt eine wichtige Rolle, doch auch Formüberlappung trägt in gradueller Weise bei und nicht als einfache Ja–Nein–Eigenschaft. Durch die Verwendung kontinuierlicher Maße phonologischer und semantischer Ähnlichkeit liefert die Dissertation eine feinere Beschreibung darüber, wie verwandte Wörter die Erkennung beeinflussen.

Über die Verhaltensanalysen hinaus bewertet die Dissertation, wie gut unterschiedliche theoretische Ansätze die beobachteten Effekte erklären können. Es werden zwei breite Klassen von Theorien betrachtet. *Distributionell-konnektionistische* Theorien verstehen Worterkennung als das Ergebnis erlernter Zuordnungen zwischen Form und Bedeutung, ohne explizite Repräsentationen von Morphemen anzunehmen. *Lokalistische* Theorien hingegen gehen von diskreten Repräsentationen für Wörter und Morpheme aus und erklären Familiengrößeneffekte häufig durch Aktivierungsausbreitung zwischen verwandten Repräsentationen. Die Dissertation implementiert computergestützte Modelle, die jeweils einen dieser Ansätze repräsentieren, und testet, wie gut sie auditiven lexikalischen Entscheidungsreaktionszeiten vorhersagen.

Auf der distributionell-konnektionistischen Seite wird eine auditive Version des *Discriminative Lexicon Model* angewendet. Dieses Modell sagt Reaktionszeiten auf Grundlage dessen vorher, wie gut akustischer Input auf semantische Repräsentationen abgebildet werden kann. Das Modell erklärt einen Teil des auditiven Familiengrößeneffekts und zeigt, dass einige

Aspekte dieses Effekts aus allgemeinen Lernmechanismen entstehen können, die geteilte Form–Bedeutungs–Strukturen nutzen, selbst ohne explizite Morphemrepräsentationen. Es erklärt den Effekt jedoch nicht vollständig.

Auf der lokalistischen Seite wird das Modell der gesprochenen Worterkennung *DIANA* verwendet. DIANA generiert und aktualisiert Hypothesen über mögliche Wörter, während sich das akustische Signal entfaltet. Aus diesen Hypothesen abgeleitete Maße erklären einen Teil des Familiengrößeneffekts, obwohl das Modell keine expliziten morphologischen Beziehungen kodiert. Dies wird darauf zurückgeführt, dass Wörter mit vielen Familienmitgliedern häufig auch viele ähnlich klingende Nachbarn haben, was den Hypothesenraum des Modells beeinflusst. Das Hinzufügen eines expliziten Aktivierungsausbreitungsmechanismus zu DIANA, der morphologische Beziehungen modellieren soll, verbessert die Leistung nicht. Insgesamt schneiden die beiden Modellierungsansätze ähnlich ab, sobald Familienmitglieder nach Form- und Bedeutungsähnlichkeit gewichtet werden. Die Dissertation kommt daher zu dem Schluss, dass die derzeitige Evidenz keinen der beiden Mechanismen eindeutig bevorzugt.

Abschließend reflektiert die Dissertation die Rolle computergestützter Modellierung in der Forschung zur gesprochenen Worterkennung. Die Implementierung von Theorien als explizite Modelle zwingt dazu, Annahmen klar zu formulieren, und zeigt auf, wo Konzepte wie Frequenz, Ähnlichkeit oder morphologische Struktur sorgfältig definiert werden müssen. Zugleich sind Modelle durch die Daten und Aufgaben begrenzt, auf die sie angewendet werden. Lexikalische Entscheidungsexperimente liefern wertvolle Einsichten, haben jedoch auch klare Grenzen. Die Dissertation plädiert daher für zukünftige Arbeiten, die Modellierung mit anderen Methoden, etwa neurophysiologischen Messungen, kombinieren, um den zeitlichen Verlauf der gesprochenen Worterkennung besser zu erfassen.

Zusammenfassend zeigt die Dissertation, dass morphologische Struktur die Erkennung gesprochener Wörter beeinflusst, jedoch in gradueller und wortabhängiger Weise. Sowohl stammbasierte als auch ganzwortbasierte Information tragen bei, und gesprochene Wörter zeigen unter bestimmten Bedingungen einen klaren Effekt der morphologischen Familiengröße. Durch die Analyse groß angelegter Verhaltensdatensätze mit expliziten computergestützten Modellen zeigt die Arbeit, dass morphologische Effekte nicht zwingend explizite Morphemrepräsentationen erfordern, sondern auch aus

allgemeineren Eigenschaften des lexikalischen Systems hervorgehen können. Insgesamt tragen diese Befunde zu einem besseren Verständnis der Erkennung morphologisch komplexer gesprochener Wörter bei und unterstreichen den Wert der Integration empirischer Daten mit formaler Modellierung.

*

Curriculum Vitae

Hanno Müller was born on 3 October 1991 in Cologne, Germany. He obtained a Bachelor of Arts (BA) in German Language and Literature from the Albert Ludwig University of Freiburg in 2016. In 2019, he completed two Master's degrees at the same institution: a Master of Arts (MA) in German Linguistics and a Master of Science (MSc) in Cognitive Science.

During his studies, he held several student research assistant positions and worked as a freelance journalist for the *Badische Zeitung*. In 2019, Müller began his PhD within the Speech Production and Comprehension group at the Centre for Language Studies at Radboud University Nijmegen, and as a member of the DFG Research Unit FOR 2373 *Spoken Morphology* at the Department of English Language and Linguistics at Heinrich Heine University Düsseldorf. During his doctoral training, he successfully completed the Education and Training Programme of the International Max Planck Research School for Psycholinguistics.

He is currently employed as an AI Engineer at the Hasso Plattner Institute in Berlin, working at the AI Service Centre funded by the German Ministry of Research, Technology and Space. His work focuses on natural language processing and speech technologies.

Alongside his academic and professional activities, Müller has been engaged in nature conservation. Since 2018, he has held an honorary position with the *Waddenvereniging* as an excursion guide in the Wadden Sea and served as financial controller and board member of the *Werkgroep Excursies* from 2019 to 2022. He is active as a wildlife and nature conservation photographer, with his work published in various print and online media.

*

Acknowledgements

Dit boek zou nooit tot stand zijn gekomen zonder de steun van allen aan wie deze dankbetuiging is opgedragen. Allereerst wil ik mijn promotor **Mirjam** bedanken. Je hebt mijn ideeën tot in de kleinste details kritisch getoetst. Als ik zou schrijven dat het niet gemakkelijk was om je te overtuigen, dan zou dat nog een understatement zijn. In verhitte discussies vlogen de argumenten ons om de oren, en zo bleven uiteindelijk alleen die theorieën overeind die het waard waren om verder te worden onderzocht. Daarmee heb je wezenlijk bijgedragen aan de kwaliteit van dit werk. Je hebt niet alleen de ondankbare rol van advocatus diaboli vervuld, maar mij ook met al je ervaring en vooruitziende blik ondersteund. Je gaf mij waardevolle ideeën voor de structurering van teksten, adviseerde mij bij het verwerken van reviews, hielp bij het plannen van werkpakketten, en nog veel, veel meer. Wat heb ik allemaal van jou geleerd!

Een groot deel van dit werk rust op een technisch veeleisend fundament dat ik zonder mijn copromotor **Louis** niet had kunnen opbouwen. Of het nu ging om inferentiële statistiek of waarschijnlijkheidstheorie, lineaire algebra of Hidden Markov Models – met welke methoden ik mij ook bezighield, je stond mij steeds met raad en daad bij. Samen werkten we ons door de aannames en voorwaarden van verschillende modelleringsbenaderingen, interpreteerden we complexe interacties, of zetten we modeltijden om in reactietijden en log-geschaalde effectgroottes in milliseconden. Of het nu ging om academische uitwisseling of gewoon even bijkletsen, voor een kopje koffie in de keuken nam je altijd graag de tijd. Als ik me eens somber voelde, ging het me na een gesprek met jou meestal weer beter.

Het project Dutch Morphological Complex Words zou zonder **Tim** niet compleet zijn geweest, en ook mijn PhD-ervaring zou zonder al onze gesprekken iets hebben gemist. We spraken niet alleen over taalkunde en statistiek, maar ook over politiek, geschiedenis en economie; over auto's en horloges, Duitsers en Nederlanders, vogels, vakantie, sport en familie. De sprint naar de negende verdieping in het Erasmusgebouw zal ik nooit vergeten. Onze conferentiereis naar Niagara-on-the-Lake was zonder twijfel een hoogtepunt van mijn PhD. You and **Patricia** gave me a helping hand in organising the final matters during the preparation period for my defence. For this, among others, I would like to thank you, my paranymphs.

My sincere thanks go to the manuscript committee – **Martha** Larson (chair), **Harald** Baayen, **Ingo** Plag, **Sabine** Arndt-Lappe, and **Stefan** Frank – who reviewed my doctoral dissertation and provided valuable comments for its improvement.

Für die Erstellung der Vogel-Zeichnungen möchte ich meiner Mutter **Klaudia** danken. Du hast dich in die Morphologie von Vogelschnäbeln, -federn, und -eiern eingearbeitet und dir die zeichnerische Darstellung der Vegetation von Salzwiesen angeeignet. Mit deinen Zeichnungen erhält diese Dissertation eine ganz persönliche Anmutung. **Jenny**, vielen Dank für das Einscannen, Layouten und Formatieren der Zeichnungen. Ohne deine Hilfe hätte ich nicht gewusst, wie ich die Zeichnungen in dieses Buch hätte integrieren können. Ik wil ook **Guus** en het team van **Proefschrift AIO** bedanken, die mijn ideeën voor de vormgeving van dit boek op prachtige wijze hebben gerealiseerd.

A big thank-you goes to **all my (PhD) colleagues**. Mijn speciale dank gaat uit naar **Saskia** voor haar hulp bij het zoeken naar woonruimte, evenals voor de daaropvolgende tuinprojecten, de vogelgespreken en een heerlijk dansje op de Fusion! **Elly** en **Thijs**, jullie kunnen je niet voorstellen hoeveel jullie gastvrijheid voor mij heeft betekend – ik weet niet wat ik zonder jullie had moeten doen. **Laia**, you also offered me a roof over my head, even though we hardly knew each other at the time – thank you very much for that! Toen ik in Nijmegen aankwam, kende ik niemand, **Marlou**; onze gezamenlijke activiteiten waren altijd een prettige manier om mijn tijd door te brengen. **Martijn**, onze korte periode als kantoorgenoten en onze koffiegesprekken waren altijd verrijkend – bedankt voor je loyaliteit als abonnee van mijn kalenders! **Elena**, many thanks for taking responsibility for my plant babies. **Chantal**, toen ik op het punt stond om de handdoek in de ring te gooien, heb je me nieuwe moed gegeven. **Chen**, thank you for a wonderfully relaxing tea ceremony. **Aurora**, **Emily**, **Figen**, **Katherine**, **Lotte**, **Wei**, and **Yu**, thank you all for the sociability during coffee and/or lunch breaks. **Theresa**, dein Sandwich-Toaster macht mich noch immer glücklich.

Henk heb ik te danken voor waardevolle opmerkingen over Research Data Management. Al mijn vragen over Ponyland, het computercluster van het Centre of Language Studies, werden snel beantwoord door **Wessel**. Voor vragen over de organisatorische aspecten van het PhD-traject vond ik altijd hulp en ondersteuning bij de **Graduate School**, vooral dankzij **Peter** en later **Suzanne**. Dat was vooral tijdens de pandemie een belangrijke steun. **Kevin**, you

stand for events that bring PhDs together, pizza after work, and conversations about photography that could go on for hours. Who else but you could be the face of the International Max Planck Research School?

A huge thank-you goes to the **Speech Production and Comprehension** research group for the open and consistently constructive exchange. With concentrated expertise on morphology and word recognition, I received valuable feedback from the **DFG Research Unit FOR 2373 *Spoken Morphology***. Der Schreibaufenthalt mit **David**, **Dominik**, **Ghattas**, **Marie**, **Simon** und **Viktoria** zählt zu den produktivsten Wochen meines Doktoratsprojekts. Die Spieleabende nach einem anstrengenden Arbeitstag waren natürlich das i-Tüpfelchen.

Motoki, deine Überraschung darüber, wie viel die Leute bei ‚Secret Identity'-Spielen lügen, bringt mich immer noch zum Schmunzeln. Vielen Dank für dein exorbitantes Engagement in unseren Nebenprojekten und dafür, dass du mich in Kneipen in Kanada, den Niederlanden und Deutschland begleitet hast!

I am grateful to the **Quantling group** in Tübingen for introducing me to Generalised Additive Mixed Models. Special thanks go to **Yu-Ying** for our joint discussions of the advantages and disadvantages of different ways of modelling random effects. **Maria,** ich möchte dir nicht nur für den regelmäßigen, motivierenden Austausch während der letzten Phase des PhD-Projekts danken, sondern auch für eine wunderbar erfrischende Wanderung im herbstlichen Frontenac Provincial Park.

Ob am Holbeinplatz mit Klapptisch vor dem Bett, ob im engen Transporter irgendwo bei Bremen, ob am See in Motzen oder ob in unserer heutigen Bleibe in Berlin – **Lu**, du hast mich beim Anfertigen dieser Doktorarbeit ‚hinter der Kulisse' begleitet. Wir haben in diesen Jahren zwei Autos zum Campen hergerichtet, eine Wohnung renoviert und eine Küche umgezogen. Wir haben unzählige Brettspielabende hinter uns, hunderte gewanderte und tausende gefahrene Kilometer, Reisen nach Südkorea, Schweden, Polen, die Niederlande, Frankreich, Spanien, Portugal und Italien. Somit bedeutete mein Promotionsstudium nicht nur hartes Arbeiten und langes Nachdenken, sondern vor allem auch Puzzeln, Abenteuer und Lust am Leben. Ich danke dir für all die verschiedenen Formen deiner Unterstützung.

9 789465 152059 >

Radboud Universiteit