Words and Paradigms in the Mental Lexicon

by

Kaidi Lõo

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics
University of Alberta

Examining Committee:

Juhani Järvikivi, Supervisor
R. Harald Baayen, Supervisor
Antti Arppe, Supervisory Committee
Herbert Colston, Examiner
James P. Blevins, Examiner

Abstract

This dissertation examines the comprehension and production of Estonian case-inflected nouns. Estonian is a morphologically complex Finno-Ugric language with 14 cases in both singular and plural for each noun. Because storing millions of forms in memory seems implausible, languages like Estonian are often taken to be prime candidates for rule-driven morpheme-based processing. However, not all Estonian nouns actually occur in all their 28 cases, but only in cases that make sense based on the meaning of the word. For instance, for *jalg* 'foot/leg', the nominative plural *jalad* 'feet/legs' is very common, whereas the essive singular case *jalana* 'as a foot/leg' rarely ever gets used. Furthermore, Estonian inflected forms cluster into inflectional paradigms, which typically come with only a few inflected variants from which other forms in the paradigm can be predicted. Hence, the number of forms that a speaker would need to memorize is much smaller than the number of forms that one can understand or produce.

Based on these observations, we aimed to clarify lexical-distributional properties that co-determine Estonian processing. Using a large number of items and generalized mixed effects modeling, we tested the influence of a number of lexical measures, such as lemma frequency, whole-word frequency, morphological family size, inflectional entropy, orthographic length and orthographic neighbourhood density (all calculated on the basis of a 15-million token Estonian corpus). Importantly, we hypothesized that a new measure, the number of attested forms of a given paradigm, i.e., the forms that actually get used, may affect morphological processing in Estonian.

We conducted four psycholinguistic experiments: two word naming tasks, a lexical decision and a semantic categorization experiment with native speakers of Estonian, varying in age (21-69 years). Results of the semantic categorization task with 200 inflected forms showed a facilitatory effect of inflectional paradigm size in both response times and accuracy. In the word naming, which had with similar number of items, a facilitatory effect of whole-word frequency was found. In the two remaining large-scale-studies, a lexical decision task and a word naming task, with over 2,000 inflected forms, both whole-word frequency and inflectional paradigm size again emerged as the strongest predictors. In line with the behavioural data, eye movement data collected during the word naming task further confirmed whole-word frequency and inflectional paradigm size as the main predictors of Estonian inflected word processing. Further analyses of pupil dilation supported these findings, but also suggested large individual differences in processing patterns. In summary, our findings suggest a surprising amount of item-specific knowledge is available during language processing. This contradicts a purely decompositional approach to the processing of complex words, even for a language as morphologically rich as Estonian.

Preface

The research projects contained within this dissertation received research ethics approval from the University of Alberta Research Ethics Board, Project Name Lexical processing of complex words", No. Pro00054925, 30 March 2015 and "Lexical processing in Estonian", No Pro00057626, 07 July 2015. The research conducted for this thesis was done in collaboration with other researchers.

Chapter 2 is accepted for publication as Lõo, K., Järvikivi, J., & Baayen, R. H. (accepted). Whole-word frequency and inflectional paradigm size facilitate Estonian case-inflected noun processing. *Cognition*. I was responsible for experiment design, data collection, analysis, and manuscript composition. Dr. R.H. Baayen and Dr. J. Järvikivi were the supervisory authors, Dr. R. H. Baayen assisted with experimental design, statistical analysis, concept formation and manuscript edits, Dr. J. Järvikivi helped with concept formation and manuscript edits.

Chapter 3 is accepted for publication as 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*. I was responsible for experiment design, data collection, analysis, and manuscript composition. Dr. R.H. Baayen and Dr. J. Järvikivi were the supervisory authors, Dr. R. H. Baayen assisted with statistical analysis, concept formation and manuscript edits, Dr. J. Järvikivi helped with experimental design, concept formation and manuscript edits. Dr. F. Tomaschek helped with experimental design and manuscript edits, Dr. J. van Rij helped with experimental design, and Dr. B.V. Tucker helped with statistical analyses and manuscript edits.

The study in Chapter 4 was conducted with the assistance from Dr. J. Järvikivi, Dr. R.H. Baayen and Dr. J. van Rij. I was responsible for experiment design, data collection, analysis, and manuscript composition. Dr. J. Järvikivi helped with concept formation and manuscript edits, Dr. R.H. Baayen helped with manuscript edits and statistical analysis, Dr. J. van Rij helped with statistical analyses.

Chapter 5 is published as Lõo, K., van Rij, Jacolien, Järvikivi, J., & Baayen, R. H. (2016). Individual differences in pupil dilation during naming task. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J. C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Austin, TX, pp. 550-555. I was responsible for experiment design, data collection, analysis, and manuscript composition. Dr. J. van Rij helped with experimental design, statistical analyses, concept formation and manuscript edits. Dr. R.H. Baayen helped with statistical analyses, concept formation and manuscript edits. Dr. J. Järvikivi helped with manuscript edits.

Acknowledgements

Funding for this dissertation was provided by the national scholarship program Kristjan Jaak, funded and managed by Archimedes Foundation in collaboration with the Estonian Ministry of Education and Research. I would also like to thank the Alexander von Humboldt foundation for support for research visits to Tübingen (AvH Professur grant nr. 1141527 to R. H. Baayen), and Words in the World Partnered Training Initiative (funded by the Social Sciences and Humanities Research Council of Canada), the GSA doctoral travel fund, Kristjan Jaak Scholarship for Short Study Visit grant and travel awards by the Department of Linguistics for funding my conference visits.

This dissertation and my studies at the Department of Linguistics of University of Alberta would have not been possible without the help of many wonderful people.

First and foremost, I want to thank my supervisors Dr. Juhani Järvikivi and Dr. Harald Baayen. Their guidance, patience and sense of humour have not only made this possible but also enjoyable. Juhani's unwavering confidence in me has helped me to grow not only professionally but also personally, and I deeply appreciate all the time he has dedicated into mentoring me. Without Harald I would not have set out to pursue a PhD. His passion for research has been inspiring and I am forever grateful for all that I have learnt from him.

My gratitude also goes to Dr. James P. Blevins for being the external examiner of my dissertation and for providing me insightful comments and questions on Estonian morphology. I would also like to thank my supervisory committee member Dr. Antti Arppe, for his questions and comments during my candidacy and defense as well as my external Dr. Herbert Colston for his questions and comments during my defense. Further, I would like to thank Dr. Benjamin V. Tucker for his help with Chapter 3 as well as Dr. Gary Libben and Dr. Pärtel Lippus for being part of my candidacy committee.

I was very fortunate to benefit from two excellent research groups during PhD. The Centre for Comparative Psycholinguistics in Edmonton has been my second (if not the first) home. First of all, I owe a massive gratitude to Abigail G. Toth, the honorary fourth member of my supervisory committee. Her friendship and support mean the world to me, and her careful proofreading and thoughtful comments made my whole dissertation much nicer to read. Also, a thanks to Dr. Caelan Marville for making my life easier by providing me a LaTex template for this dissertation. Finally, I would also like to thank all the lovely people at CCP and at the Department of Linguistics for helping me to stay sane through the entire process of dissertation writing. Among others, I am thankful for Anja, Bekka, Brian, Figen, Filip, Graham, Hannah, Izzy, Keely, Lauren, Lena, Lila, Nik, Stephanie, Vince and Yoichi for that.

I also had the pleasure to go back to the Quantitative Linguistics Lab in Tübingen multiple times during my PhD. The innovative and stimulating environment there during my Masters was something which got me interested in psycholinguistics in the first place. I would like to express my deepest thanks to Dr. Jacolien van Rij and Dr. Fabian Tomaschek, who patiently helped me through the many hurdles of experimental research. I would also like to thank Arvi, Cyrus, Petar, Tineke and Tino for making my life easier in various aspects during my visits in Tübingen.

Furthermore, I would like to thank Dr. Heiki-Jaan Kaalep and Dr. Fermin Moscoso del Prado Martin for their help with my corpus measures. I am also grateful to Dr. Pärtel Lippus, Dr. Aki Kyöroläinen, Dr. Einar Meister and Dr. Mari Uusküla for their help with the administration of my experiments in Estonia, and to all the wonderful participants who took part in my experiments.

Lastly, I wish to dedicate this dissertation to my dear family and friends who have been there for me throughout this PhD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We hear and speak hundreds, maybe thousands of different words every day. Most of the time this process is effortless and goes unnoticed. One of the goal of psycholinguistic research is to understand what makes these processes possible, and more specifically, what factors and mechanisms underlie human word comprehension and production. This question is especially pertinent for languages like Estonian that has highly productive morphological processes, such as compounding (e.g., *jalgpall* 'football'), derivation (e.g., *jalgsi* 'afoot') and particularly inflection (e.g., *jalad* 'feet'), which all create new words and word forms. This dissertation will concentrate on the processing of Estonian case-inflected nouns.

95% of word forms encountered daily by Estonian speakers are morphologically complex. Estonian is particularly rich with respect to inflection, with 14 inflectional cases in singular and plural, in addition to parallel (or overabundant) forms and rich compounding. This increases the number of possible forms to millions. Unlike English, which uses prepositions, Estonian makes use of case marking. A three-word prepositional phrase in English, such as *on the feet*, is expressed by using only a single word in Estonian, *jalgadel*. The Word Atlas of Language Structures (WALS) [1] entailing information on case marking of 261 different languages, lists Estonian together with 24 other languages which have 10 or more grammatical cases. More precisely, only a few other languages in the database, primarily other Finno-Ugric languages, like Finnish (15 cases), Udmurt (16 cases) and Hungarian (18 cases), which have a richer inflectional systems than Estonian.

[1] http://wals.info

Because storing an abundance of forms in memory may seem implausible, as argued by a number of researchers (Hankamer, 1989; Niemi, Laine & Tuominen, 1994; Yang, 2016), it is often assumed that languages like Estonian must undergo rule-driven morpheme-based processing. According to this view, all complex words are broken down into morphemic components, which are stored in the mental lexicon. However, a growing body of research challenges this assumption. It has been found that both properties of individual forms, such as whole-word frequency, as well as properties of paradigmatic relations, such as morphological family size (i.e., the number of complex derived and compound words sharing a constituent) and inflectional entropy (i.e., the average amount of information within a paradigm), seem to take precedence of morpheme-based information (see e.g., Baayen, Dijkstra & Schreuder 1997; Milin, Kuperman, Kostić & Baayen 2009; Schmidtke, Matsuki & Kuperman 2017). As the previous research has primarily investigated languages with impoverished morphological systems, the natural question is to ask, to what extent do speakers of morphologically rich languages, such as Estonian, benefit from these factors?

Interestingly, despite the large number of inflected forms available for each Estonian word, not all forms are used equally as often or at all. Thus, the number of possible forms actually used in Estonian can be much smaller than theoretically possible. In other words, if we look at a representative corpus of Estonian, we notice that not all Estonian nouns occur in all 14 possible cases, but only in cases which make sense semantically. For instance, the nominative plural *jalad* ('feet') occurs quite frequently, however, the essive singular case *jalana* ('as a foot') rarely ever occurs, as it does not make much sense semantically.

The inflected variants available for a particular word make up its inflectional paradigms. Based on the findings from morphological family size and inflectional entropy, it is interesting to ask whether the number of forms that get used, i.e., a word's actual *inflectional paradigm size*, as opposed to the theoretically possible inflectional paradigm size, might have an effect on lexical processing.

The purpose of the current dissertation is to shed light on the level of linguistic granularity and mechanisms necessary for the processing of Estonian case-inflected nouns. With its morphological complexity and the varying size of actually attested inflectional paradigm members, Estonian offers an interesting case for the investigation of language processing that would not be possible with morphologically less rich languages. To do this, the current dissertation takes a multi-methodological approach, applying eye tracking (both eye movement and pupil size measures), lexical decision, semantic categorization and word naming to capture different aspects of inflectional morphology affect how Estonian native speakers from diverse age range comprehend and produce case-inflected nouns. The dissertation aims to expand our knowledge on 1) the properties of individual inflected forms and their inflectional paradigms influence visual word recognition (i.e., reader's ability to read and recognize written isolated words) and word production (i.e., reader's ability to articulate written isolated words); 2) the time course of inflectional processing; and 3) the role of individual differences in inflectional processing.

In what follows, I will give an overview of both linguistic and psycholinguistic approaches to morphologically complex words, followed by how Estonian inflection work. After that, I will give a summary of the previous research reporting whole-word and paradigmatic effects in morphological processing. I will then provide a synopsis of the current dissertation, including an overview of the three main research questions as well as the experimental and statistical methods used. Finally, I will provide a summary of the individual studies that form the main body of the current dissertation.

1.1 Linguistic approaches to complex words

Words differ in their structure, their meaning and the way they are combined with one another. The linguistic field of morphology is interested in the systematic correspondences between form and meaning. Different approaches have been proposed to describe and explain how these mappings come about. In the literature, three general frameworks stand out: Item and Arrangement, Item and Process, and Word and Paradigm morphology.

According to the **Item and Arrangement** approach, a term coined by Hockett (1954), all complex words are composed of word subunits called morphemes. Morphemes are arranged linearly in a certain fixed order and each morpheme has its own meaning and function. A complex word is analysed as the sequence of these subunits, for example, the meaning of *cats* is a composition of the meaning *domestic feline mammal* and the meaning *plurality*.

A classical representative of the Item and Arrangement theory is the Word Syntax approach suggested by Di Sciullo & Williams (1987); Lieber (1981, 1992) and Selkirk (1982). According to their approach, both affix and nonaffix morphemes have an identical lexical entry. However, affixes entail subcategorisation frames, which specify the syntactic category (e.g., verb, noun, adjective) of the affix can attach to the syntactic category that the combination produces. For example, the subcategorization frame for English affix *-ness* specifies that it can be attached to an adjective (e.g., *dark*) and that the resulting word is a noun (e.g., *darkness*).

This approach fits well with fully agglutinative languages, such as Turkish, where complex words almost always consist of clearly distinguishable parts. In such languages, it is seemingly easy to assign a function to each subunit. Item and Arrangement may be also good at describing isolating languages, such as Ancient Chinese, where each word is a single morpheme. However, Item and Arrangement approach is challenged by languages where morphological formatives are less clearly distinguishable bundles of semantic and syntactic features, having often more than one meaning (also called portmanteau morphs). For example, in the Estonian partitive plural *kalu* ('fish'), the formative *-u* denotes both the partitive case and plurality.

The second theory, **Item and Process**, a term also coined by Hockett (1954), is less concerned with morphemes, but rather the main focus is on operations that build complex words. Word formation is characterized by applying specialized rules to the base form (i.e., the simplest form). For instance, English plurals are formed by a rule that states "add *-s* to the base form "(e.g., from *house+s* to *houses*), however, irregular plurals are formed by applying an ablaut rule to the base form *man* (e.g., from *man* to *men*).

A representative of the Item and Process approach can be found in the works of Aronoff (1976, 1992, 1994), according to his theory rules are an essential part of the lexicon, however, the main units on which rules operate are not morphemes, but lexemes. Aronoff (1992) defines a lexeme as a member of a major lexical category (either noun, verb, adjective or adverb) that has both form and meaning, but can exist outside any particular syntactic context, i.e., morphology can exist by itself. For instance, the English past participle can be created by three different processes: suffixation (e.g. *taken*), ablaut (e.g., *sung*), or the combination of the two (e.g., *broken*). Hence, one cannot represent the English past participle with a particular suffix morpheme, rather it exists as an abstract category.

Distributed Morphology (Bobaljik, 2012; Embick & Noyer, 2007; Halle & Marantz, 1993; Harley & Noyer, 1999) is a theory midway between Item and Arrangement and Item and Process approaches. It is similar to the traditional Item and Arrangement theory in that it models words by means of hierarchical structures of morphemes, however, morphemes in this approach are considered atomic abstract (form-free) units, like *plurality* or *past tense*, which exist only on a syntactic level. Therefore, *words* and *the lexicon* are not important concepts and as a consequence, paradigms are also not important. Phonological forms are realized locally by spell-out rules in the syntactic tree, which have access to the lists of semantic and sometimes syntactic features. Morphemes figure in syntactic structures as inputs for compositional semantic rules. It is important to note that although Distributed Morphology is not a pure IA theory, it is possibly the only major currently applied morphological theory that considers the morpheme as a necessary discrete mental unit. Hence, Distributed Morphology seems to be compatible with the sublexical approach to the psycholinguistics of complex words, introduced below in Section 1.3.

Paradigm Function morphology (Stewart & Stump, 2007; Stump, 2001, 1991, 1993a,b) and A-Morphous morphology (Anderson, 1977, 1982, 1992) make use of a realizational theory. Realization rules operate on a lexical root and on a bundle of morphosyntactic features that needs to be realized. The derivation of a word form occurs by means of the paradigmatic class it belongs to. Word formation is thus amorphous: the realization rules are simply pairings of phonological representation and its features. For example, the German lexical base <*Buch-,{genitive','singular'}*> feature pairing gets realized in the morphotactically unstructured surface form *Buches* 'genitive singular book' when it occupies the genitive singular cell in the correct paradigm class. Realization rules are organized into blocks; rules that fill in the same slot in the paradigm are placed in one rule block. In the situation where more than one rule applies, the more specific rule takes precedence. For instance, the plural form *oxen* can be derived from *ox* because the irregular plural rule precedes the regular rule, which would generate the incorrect form **oxes*. In computational morphology, Karttunen (2003) shows that realizational morphology is in fact equivalent to a finite state morphology.

Thus, Item and Process theory has moved away from the traditional Bloomfieldian notion of morpheme as a sign in the sense of the Saussurean sign. This is also reflected in terminology as the followers of this approach discuss exponents and formatives rather than morphemes. Finally, Word and Paradigm, the last approach to morphological theory, has moved away from morphemes completely.

Word and Paradigm morphology (Matthews, 1974; Robins, 1959) takes the focus away from the parts of the word and reallocates the attention to the word as a whole. As the name implies, words are organized in paradigms, where new forms are created on the basis of proportional analogy across paradigms. Any of the members can be reconstructed based on other forms, once the similarities and differences in analogy are recognized. For instance, in the four-part equation *walk*:*walked* is equal to *talk:x*, by analogy *x* equals *talked*. The general idea is that such analogical patterns are also expected to hold for much larger paradigms. Thus, words in Word and Paradigm morphology do not exist in isolation like in the previous two theories, but are a part of a larger system.

Current developments within Word and Paradigm theory are mainly lead by James P. Blevins (Blevins, 2003, 2006, 2013, 2016). In the recent work, he formalized the core concepts of analogy in terms of information theory by estimating the entropy of paradigm cells. He argues that depending on the paradigm, some cells are more uncertain than others, i.e., they have a higher entropy. Blevins notes that for Estonian, when a paradigms contains a strong vowel-final partitive singular, such as *lukku* 'lock', the rest of the singular forms are predicted. For example, the nominative singular *lukk* is just the partitive singular without the stem vowel *-u*; and the genitive singular *luku* is just the partitive singular in the weak grade[2]. Finally, all locative cases follow from the genitive singular stem and the appropriate case marker (e.g. the illative singular *luku-sse* ' into the lock', the inessive singular *luku-s* 'in the lock/locked', the elative singular *luku-st* ' from the lock'). Moreover, paradigm cells are not independent from each other, as most of the singular forms in the *lukk*-paradigm can be also used to predict each other, as well the partitive case. As a result, each form in the paradigm is highly informative, i.e., has a low entropy, and the entropy of the whole paradigm is lower than the summed entropy of the individual entropies.

A different approach to analogy is taken by the machine learning approach of the Tilburg Memory Based Learner (Daelemans, Zavrel, Van der Sloot & Van den Bosch, 2007). Computationally, this model takes as input sets of feature values and their associated class and constructs a nearest-neighbourhood based algorithm to classify new input. For instance, the model is able to predict the correct class of a Dutch diminutives: either ending with *-etje*, *-tje*, *-je*, *-kje*, or *-pje*. This is done based on a detailed phonological feature set which specifies whether the last syllables of the input are stressed, have onsets, nucleae or codas.

As described above, linguistic theories have handled morphologically complex words in many diverse ways, which include analogy, realization rules, subcategorisation frames and structure trees. With respect to the current research, this raises two questions: first, how informative are these theories with respect to processing in Estonian? And second, what is relationship between these theories and psycholinguistic theories of morphological processing? I.e., how informative are they with respect to psycholinguistic approaches to complex words? Before we turn to psycholinguistic models of complex words, I will give a brief introduction of Estonian nominal inflection. The description and examples below are primar-

[2] Consonant gradation is a an alternation typical for Finno-Ugric languages in which certain consonants alternate between short, long and (in some languages) overlong length.

ily based on two sources, *Eesti keele käsiraamat* ('Handbook of Estonian language') (Erelt, Erelt & Ross, 2007) and *Estonian language* (Erelt, 2003).

1.2 Estonian nominal inflection

Inflected forms in Estonian can be created in two different ways. First, inflection can be agglutinative – stems and affixes can be clearly identifiable as stems and exponents that realize functions, such as number and locative case. For example, in the inessive plural *kalades* 'in the fishes', *kala* is the stem, *-de* marks plurality and *-s* marks the locative case inessive. Second, Estonian also makes use of fusional patterns, where it is often unclear whether a complete form is a stem alternate or if it includes a stem and a portmanteau morph. For example, one may take the whole form *kalu* as the stem alternate, or alternatively, one may distinguish the stem *kal* and a portmanteau morph *-u* which stands for the partitive plural.

Table 1.1: Inflectional paradigm of kala 'fish' with 29 members.

Case	Singular	Plural	English translation
Nominative	kala	kalad	fish (subject)
Genitive	kala	kalade	of a fish/fish (total object)
Partitive	kala	kalasid, kalu	fish (partial object)
Illative	kalasse	kaladesse	into a fish
Inessive	kalas	kalades	in a fish
Elative	kalast	kaladest	from a fish
Allative	kalale	kaladele	onto a fish
Adessive	kalal	kaladel	on a fish
Ablative	kalalt	kaladelt	from a fish
Translative	kalaks	kaladeks	[to turn] into a fish
Terminative	kalani	kaladeni	up to a fish
Essive	kalana	kaladena	as a fish
Abessive	kalata	kaladeta	without a fish
Comitative	kalaga	kaladega	with a fish

A typical Estonian inflectional paradigm consists of 14 inflected forms in both the singular and plural (see Table 1.1; adapted from Erelt et al. 2007). Nominatives, genitives and partitives are considered grammatical cases, as they are primarily used to fulfil various syntactic functions.

The nominative is used to mark the subject as in ***Lind** istub pesal* '**The bird** is sitting on the nest'. The genitive typically expresses possession as in ***Linnu** pesa* '**The bird's** nest' or total object as in *Viisime **linnu** välja* 'We took **the bird** out'.

The partitive denotes partial object as in *Toitsin **lindu*** 'I was feeding **a bird** (but did not finish)'.[3] Illative, inessive, elative, allative, adessive and ablative are locative cases that primarily express spatial relations, however, they can also express time, manner and state. For example, the adessive expresses location in space or time as in *Raamat on **laual*** 'The book is **on the table**' or *Pidu on **reedel*** 'The party is **on Friday**'. The adessive can also mark manner as in *Lind on **hirmul*** 'The bird is **scared**'.

The remaining classes are usually classified as semantic cases. The translative expresses state or purpose as in *See vesi on **joomiseks*** 'This water is **for drinking**'. The terminative expresses temporal and location boundaries as in *Ta töötab **reedeni*** 'She travels **until Friday**'). The essive also indicates state or purpose as in *Ta magab **haigena voodis*** 'He is **sick** at bed'. The abessive case expresses the absence of manner or instrument as in *Ta sööb **kahvlita*** 'She eats **without the fork**'. And finally, the comitative case denotes the presence of manner or instrument as in *Ta sööb **kahvliga*** 'She eats **with the fork**'.

The large number of agglutinative and fusional forms give rise to several kinds of complexities. First, one word form can belong to several paradigms. For example, the form *mees* 'honey/man' can belong to *mesi* 'honey' paradigm, when we consider *mee* as the stem for *honey* and *-s* as the inessive singular case marker, however, it can also belong to the *mees* 'man' paradigm when we consider the whole form *mees* to be the stem for *man*.

Second, Estonian often makes use of several alternative forms available for the same paradigm cell, in particular in the plural. These forms are not always in free variance, instead they can sometimes express subtle differences in their meaning. For example, *jalgadel*, *jalul* and *jalgel* 'on the feet' all express the adessive plural, but have subtle differences in meaning. However, in many other cases parallel forms

[3] Estonian grammar makes the distinction between partial and total object, the partial object expresses imperfective activity and the total object perfective activity.

have similar meanings, for example, *kalasid* and *kalu* both express the object case of the partitive plural of 'fish'. As a consequence of alternative forms, the number of forms in inflectional paradigms can be much larger than 28. For instance, the paradigm of *jalg* 'foot' has 46 members in total (see Table 1.2; adapted from Erelt 2003). Finally, the same word form can serve different functions (syncretism), for example, the word form *kala* 'fish' serves as the nominative, genitive and partitive singular case of 'fish' (see Table 1.1).

All the forms that one must know in order to construct the inflectional paradigm are called the principal parts. Traditional grammars of Estonian point to the nominative, genitive and partitive singulars as principal parts. However, in many cases knowing the genitive and partitive singulars, or even the partitive singular alone is enough, as the nominative can be derived from the partitive singular. All singular cases, except the nominative and the partitive singulars. are constructed from the genitive singular by adding the appropriate inflectional ending. The nominative plural is constructed from the genitive singular by adding the plural marker, the genitive plural is constructed from the nominative singular, the partitive plural is constructed from the partitive singular, and the remaining plural cases are generated from the genitive plural. However, other cases besides principle parts can also be used to construct forms in the paradigm based on analogy.

As Table 1.2 illustrates, there are often alternative plural forms, for example the stem plural and *-e* plural. The stem plural is formed by changing the stem vowel, for example, the partive plural *jalgu* 'foot/leg' is created by changing the stem vowel in the strong grade from *a* to *u* (i.e., *jalga* to *jalgu*). The *-e* plural, as in the genitive plural *jalge* 'foot/leg', is rare and has evolved from the regular *-de* plural *jalgade* through fusion.

The Estonian inflectional system presents challenges to the Item and Arrangement approach. First, Item and Arrangement cannot straightforwardly handle the fusional forms of Estonian, as one could consider the whole form as either the stem alternate or the *-u* affix as both the partitive and plural marker. Second, for many forms there is no direct parallelism between form and meaning. For example, in Estonian, the partitive singular is used to generate plural locative cases, however,

Table 1.2: Inflectional paradigm of jalg 'foot' with 46 members.

Case	Singular	Plural
Nominative	jalg	jalad
Genitive	jala	jalgade, jalge
Partitive	jalga	jalgasid, jalgu
Illative-1	jalga	-
Illative-2	jalasse	jaladesse, jalusse, jalgesse
Inessive	jalas	jalgades, jalus, jalges
Elative	jalast	jalgadest, jalust, jalgest
Allative	jalale	jalgadele, jalule, jalgele
Adessive	jalal	jalgadel, jalul, jalgel
Ablative	jalalt	jalgadelt, jalult, jalgelt
Translative	jalaks	jaladeks, jaluks, jalgeks
Terminative	jalani	jalgadeni, jalgeni
Essive	jalana	jalgadena
Abessive	jalata	jalgadeta
Comitative	jalaga	jalgadega

the meaning of these plurals is not in the least related to the object meaning of the partitive, for example, the partitive singular *jalga* 'foot' does not contribute to the plural adessive meaning of *jalgadel* 'on the feet'.

The item and Process approach faces the challenge of setting up proper declension classes. The number of classes proposed in Estonian grammars ranges from 12 in *Eesti keele grammatika I* (Erelt, Erelt, Viks, Kasik, Metslang, Rajandi, Ross, Saari, Tael & Vare, 1995) to 26 in *Õigekeelsussõnaraamat* (Erelt, Leemets, Mäearu & Raadik, 2013). In line with the Item and Process approach, Kaalep (1997) built a morphological analyzer and synthesizer, ESTMORF which follows a realizational approach in that formatives remain unstructured. For instance, the adessive plural *jalgadel* 'on the feet' is derived through a set of rules which specify that the correct stem alternate is *jalga*, the correct formative is *-del*, and that the class for this form is 22. ESTMORF follows the classification by Viks (1992) with 26 declension classes.

A Word and Paradigm approach to Estonian is offered by Blevins (2005, 2008), who also works with declension classes but only arrives at 4. These classes are based on two features. The first feature is the shape of the partitive singular, specifying whether the partitive ending is (i) vocalic as in the partive singular *kala* 'fish' or (ii) *-d*/*-t* as in the partitive singular *mõtet* 'thought'. The second feature is the prosodic structure of the genitive singular, specifying whether the genitive singular has (i) a trochaic foot as in the genitive singular *al.gu.se* 'beginning' or (ii) a nontrochaic foot as in the genitive singular *puu* 'tree'. Thus, these declension classes, as determined by diagnostic contrasts in form and prosody, are purely morphophonological, and neither semantic nor syntactic.

Blevins argues for an implicational rather than derivational inflectional system for Estonian, in which the relationships between grammatical and semantic cases are symmetrical. All singular semantic cases can be created based on the genitive singular by adding the relevant affix. For example, the singular adessive *jalal* 'foot/leg' can be created by adding the ending *-l* to the genitive form *jala*. However, the adessive singular *jalal* in turn also implies that the genitive form is *jala*, because the adessive case ending is always *-l*. The same analogy applies to the relationship between the genitive, as well as between other semantic cases and between the semantic cases themselves. The endings of semantic cases as opposed to grammatical cases are invariant, that is, they always have the same ending. Furthermore, in some cases the symmetric relationship is also present between the grammatical cases. For example, the disyllabic partitive singular in the strong grade with a vowel ending, such as *lippu* 'flag', implies the weak genitive singular form *lipu*, in turn, the weak grade of genitive implies the strong grade of partitive. Finally, the same bidirectional relationship can also be held between the genitive plural and plural semantic cases. Knowing one form in a paradigm has implications with respect to many other forms in the same paradigm. Thus, in particular the Word and Paradigm approach seems to be in line with the inflectional system of Estonian.

1.3 Psycholinguistic approaches to complex words

Whereas theoretical morphology has in many respects moved away from the classical notion of morpheme, work in psycholinguistics is still very much grounded in early Item and Arrangement theories and the Bloomfieldian notion of the morpheme.

Different approaches have focused on the following questions: Is morphologi-

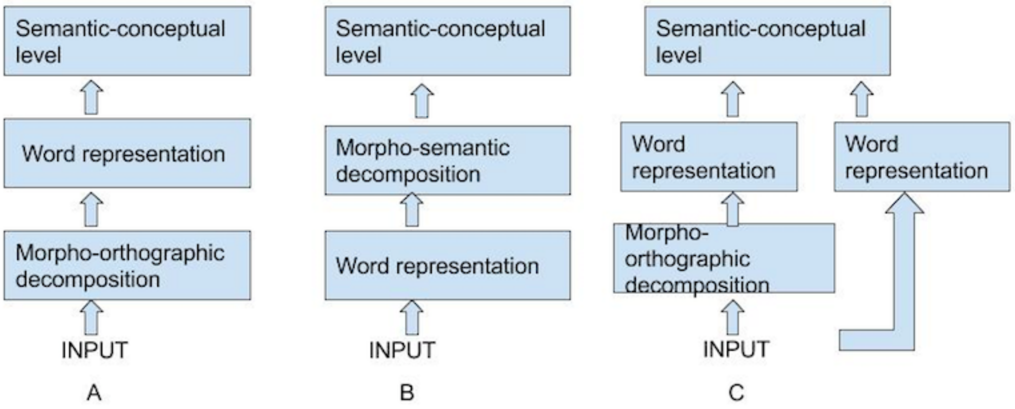
Semantic-conceptual level
Word representation
Morpho-orthographic decomposition
INPUT
A
Semantic-conceptual level
Morpho-semantic decomposition
Word representation
INPUT
B
Semantic-conceptual level
Word representation
Word representation
Morpho-orthographic decomposition
INPUT
C

Figure 1.1: Examples of the architecture of traditional models of morphological processing, which all include morphemic decomposition and whole word representations, as represented by a series of arrows and boxes.

cal structure encoded in the mind? Is morphological structure used in word recognition and production? In other words, do we decompose complex words into morphemic components when we encounter these forms? If we do in fact segment complex words into morphemes, how and when do we do it? More specifically, are all complex words decomposed or only some? And do these processes happen at the beginning of the word recognition or after the semantics of the word has been accessed?

Interestingly, the issue of whether and how morphemic decomposition would even benefit semantic composition is rarely discussed. Studies have primarily focused on the issue of when and under which conditions this decomposition occurs. Only more recently has it been suggested that the processing of complex words with respect to both storage and processing may be more plausible when complex words are immediately associated with their semantic-conceptual level, without the need for morphemic segmentation. However, before introducing these alternative approaches in more detail, I will give a brief overview of standard psycholinguistic approaches to morphologically complex words.

The **sublexical** (”form-then-meaning”) morphemic decomposition approach argues that in word recognition all morphologically complex words are parsed into morphemes before any meaning gets associated with the whole word or its parts.

This approach has received recent support from masked priming experiments, which show that transparent words such as, *cat-s*, and pseudo-complex derivational words such as *corn-er*, are both initially processed in an identical decompositional fashion (see e.g., Beyersmann, Ziegler, Castles, Coltheart, Kezilas & Grainger 2016; Kazanina 2011; Lázaro, Illera & Sainz 2016; Longtin, Segui & Hallé 2003; Marslen-Wilson, Bozic & Randall 2008; Rastle & Davis 2008; Rastle, Davis, Marslen-Wilson & Tyler 2000; Rastle, Davis & New 2004). Once obligatory morphemic decomposition has occurred, morphemic parts point to the meaning components, which will then trigger lexical access, and finally the semantics of the complex word is derived (see model A in Figure 1.1). For example, a form like *cats* is first split into morphemes *cat* and *-s*, then the mental representation of *cats* is accessed, and finally the morpheme *cat* gets directly associated with the meaning *domestic feline mammal*, and the morpheme *-s* with *plurality*. This framework has a long tradition and is still widely applied well beyond the masked priming paradigm (see e.g., Allen & Badecker 1999; Christianson, Johnson & Rayner 2005; Duñabeitia, Perea & Carreiras 2007; Fruchter & Marantz 2015; Marantz 2013; Rastle & Davis 2008; Solomyak & Marantz 2010; Taft 1994, 2004; Taft & Forster 1975, 1976a).

Another approach, the **supralexical model** (Giraudo & Grainger, 2000, 2001) was proposed to explain experimental findings that could not be accounted by the sublexical model. The supralexical model assumes initial full form lexical access, which is followed by morphemic decomposition of the form, and then, the semantics of the complex word finally gets accessed (see model B in Figure 1.1). In this approach, the word recognition of *cats* is predicted to take place the following way: first, the orthographic representation of the form *cats* is accessed, then *cats* is split into morphemes *cat* and *-s*, and finally, the morpheme *cat* gets linked with the meaning *domestic feline mammal* and the morpheme *-s* with the meaning *plurality*. Support for the supralexical model comes from a French masked priming study by Giraudo & Grainger (2001) roots and suffixed derivations provided identical priming costs. For instance, it did not make a difference whether the target *balayage* 'sweeping' was primed with the root prime *balai* 'sweep' or with the derived prime *balayeur* 'sweeper'. According to the proponents of the supralexical account, this finding has been taken to contrast with sublexical models, which predict extra processing costs for derived words, as they require decomposition. Further evidence against early decomposition is provided by Feldman and colleagues (Feldman, Pastizzo, Soltano & Francis, Feldman et al.; Feldman, 2000; Feldman, O'Connor & Moscoso del Prado Martin, 2009; Feldman & Soltano, 1999), who in a series of priming studies have shown that semantics of the whole complex word becomes relevant very early on. For instance, semantically transparent derivational relatives, such as *departure*, facilitated targets *depart*, whereas semantically opaque relatives, such as *department*, did not.

Additionally, several **dual-route models** have been proposed, such as Augmented Addressed Morphology (Burani & Caramazza, 1987; Caramazza, Laudanna & Romani, 1988; Chialant & Caramazza, 1995) and Morphological Race Model (Frauenfelder & Schreuder, 1992; Schreuder & Baayen, 1995). Dual-route models posit two processing routes, a decompositional route and a whole-word processing route, for which there is a direct connection from the orthographic representation to the meaning (see model C in Figure 1.1). For all complex words both routes compete with one another, and which route is selected depends on statistical properties, such as the stem and whole-word frequency ratio (Morphological Race Model) or on whether the complex word is already in the mental lexicon (Augmented Addressed Morphology). Support for the whole-word route comes from the experiments that show fully decomposable forms are not decomposed. For example, it has been demonstrated that plural dominant noun forms (e.g., *eyes*) are processed faster than singular dominant nouns (e.g., *noses*) even when controlled for stem frequency (Baayen et al., 1997; Baayen, McQueen, Dijkstra & Schreuder, 2003; Baayen, Schreuder, De Jong & Krott, 2002).

The **dual-mechanism model** ("words and rules") supports conditional morphological decomposition, which depends on the regularity of the complex word (Clahsen, 1999; Clahsen, Sonnenstuhl & Blevins, 2003; Hahne, Mueller & Clahsen, 2006; Marcus, Brinkman, Clahsen, Wiese & Pinker, 1995; Pinker, 1999; Ullman, 2001). According to this approach, all complex words are divided into two groups: regulars and irregulars. These two forms are assumed to be processed in fundamentally different ways, using procedural memory for regular forms and declarative memory for irregular forms. For example, the regular past in *walked* is constructed by accessing the form *walk* from the mental lexicon and by adding a suffix *-ed* to it, while the irregular past such as *sat*, is memorized and accessed directly from the mental lexicon.

Unlike in the sublexical approach, decomposition is not form-based across all complex forms, and unlike in the dual-route models, there is no competition: regular complex forms are never stored but always decomposed and irregulars in turn are always stored and never decomposed. Two types of evidence have been found to support this theory. The first comes from differential patterns of language acquisition, for example, when children are first acquiring language, they tend to overregularize irregular past tense forms, so instead of saying *sat*, they say something like **sitted*. They produce the incorrect regular form until they have had enough encounters with the correct irregular form are able to memorize it. Regular forms, such as *walked* are always produced by a rule stating ”add -ed”, thus, only the stem *walk* must memorized. The second type of evidence comes from aphasics who have difficulties with either regulars or irregulars, depending on which area of their brain is damaged. However, this view has been also extensively challenged starting with Rumelhart & McClelland (1986a)’s pioneering connectionist model (see also Joanisse & Seidenberg 1999; McClelland & Patterson 2002a,b; Ramscar 2002). With the help of computational modeling, these studies show that processing can be achieved with a singe system, grounded in semantic correspondence between phonology and semantics. This approach denies the existence of separable morphemes and argues that both irregular and regular inflected forms are learned and represented as overlapping whole forms sharing certain semantic and phonological similarities.

Moreover, the empirical evidence for a dual-mechanism approach comes almost exclusively from English. As pointed out by Blevins (2006), English is an unusually simple language when it comes to inflectional morphology. And an inflectional system as complex as Estonian poses several challenges to the dual mechanism principle of this theory. As described above, one often cannot predict Estonian inflected forms based on the inventory of stems and affixes alone. The combinatorics emerge rather automatically based on the analogy with other forms in the paradigm, rather than through rules. Further, many forms in Estonian are fusional (e.g. the partitive plural *kalu* 'fish'), where stems and affixes are not clearly separable. Following from that, one could perhaps argue that all these problematic forms are stored as irregulars. However, this is up for debate. First, memorization of inflected forms would not help to solve the issue that Estonian word formation depends on other forms in the paradigm as dual-mechanism specifies no relations to other whole-words. Second, the total number of inflected forms (including the fusional forms) that one would then need to memorize in Estonian would be very large. As argued by Hankamer (1989), we might run into the issue of exceeding the memory-capacity of the mental lexicon. Other languages run into further problems, for instance, in German, there is no one default regular plural marker such as *-s* in English, but multiple default markers (see Dabrowska 2004, for an extensive criticism).

All of the model mentioned above belong more or less to the Item and Arrangement framework, connectionist models, however, are quite different. In these models, the orthographic or phonetic input is directly associated with the conceptual-semantic level, without morphemic decomposition. Hence, there is also no difference in the processing mechanism of simple and complex words. Furthermore, these models are computationally implemented learning models: the connections between input and output are in a constant change due to learning from the context.

Parallel Distributed Processing (PDP) models were developed to simulate general cognitive mechanisms, such as problem solving and decision making (see e.g. Rumelhart & McClelland 1986b for an overview), however, there are also more specific implementations that are concerned with language. Parallel Distributed Processing uses artificial neural networks, which offer mathematically simplified approaches to learning. In distributed processing, the meaning of word is represented as a pattern of activity over many units (i.e., basic information structures, such as neurons), the meaning of a single unit cannot be interpreted alone without knowing the state of other units. Regarding morphologically complex words, the pattern of activation over hidden units represents the regularities between the form and the meaning. PDP has been successful at modelling the semantic transparency effect. Gonnerman, Seidenberg & Andersen (2007) showed that in a highly related prime and target pair, such as *baker-bake*, the priming effect was bigger than in a moderately related pair, such as *dresser-dress*, which in turn were primed more than a semantically unrelated pair, such as *corner-corn*. Plaut & Gonnerman (2000) observed increased priming effects for semantically transparent pairs in computational simulations with both morphologically rich and poor artificial languages. Both studies explain their finding in terms of distributed representations rather than discrete morphemes, hence challenging blind automatic approaches to morphological processing.

The **Naive Discriminative Learning** (NDL) model (Arnold, Tomaschek, Lopez, Sering & Baayen, 2017; Baayen, 2011; Baayen, Hendrix & Ramscar, 2013; Baayen, Milin, Filipovic Durdjevic, Hendrix & Marelli, 2011; Baayen, Sering, Shaoul & Milin, 2017) is a localist connectionist model, where input and output units are directly mapped and each unit by itself is connected to a meaning. NDL is grounded in information and learning theory (Shannon, 1948; Wagner & Rescorla, 1972). The learning algorithm specifies a set of equations that defines the association strength of n-grams (the form split into unigrams or bigrams) to some meaning (both lexical and grammatical). For example, in the Estonian plural inessive form *kalades* 'in the fishes', the cues can be *#k*, *ka*, *al*, *la*, *ad*, *de*, *es* and *s#*; and the outcomes *fish*, *plurality*, and *internal location*.

The associations between form and meaning are acquired through the experience with the language, and they are in constant change with every new input. This kind of system seems to allow for a high level of flexibility in order to deal with a variety of different morphological forms. For Estonian this means that agglutinative and fusional forms can be handled by the same system based on cue-outcome association weights. The *plurality* in the agglutinative form like the inessive plural *kalades* 'in the fishes' might be learned by associating the bigram *de* with the *plural meaning*. In a fusional form like *kalu* 'fish', the *u* might get associated with both the plural and the partitive meaning.

Interestingly, without including morphemic decomposition in the architecture, Baayen et al. (2011) showed that NDL can successfully simulate the finding in Rastle et al. (2004)'s study, one of the cornerstones of sublexical account, which showed that regular complex words (e.g., *talker*) and pseudo-complex words (e.g., *corner*) prime the stems in a similar way. Further, NDL has been used to model a series of morphological processing effects that cannot be straightforwardly explained with decompositional approaches. One example of this is the morphological family size effect (see e.g., De Jong 2002; Schreuder & Baayen 1997). Previous research has shown that a word such as *work*, which occurs as a constituent in numerous compounds and derived words (e.g., *worker*, *homework*, *work load*, *work force*), is easier to process than a word such as *ghost*, which only occurs in a few words (e.g., *ghosthunter*, *ghostly*). Using NDL, Mulder, Dijkstra, Schreuder & Baayen (2014) were able to successfully simulate the morphological family size effect with Dutch mono- and bilinguals. Further, NDL has been able to accurately predict inhibition resulting from relative entropy in Serbian (Baayen et al., 2011). Milin et al. (2009) found that the more the usage of a certain paradigm deviates from the usage of other paradigms in the same inflectional class, the less prototypical the paradigm is, and the higher the processing costs. This can be quantified with a relative entropy measure, which takes into account both the probabilistic characteristics of individual variants in the paradigm, as well as the inflectional class it belongs to. Other simulations by NDL include modeling word sequence and idiom frequency effects in English, and anti-frequency effects in Vietnamese (see e.g., Baayen et al. 2013; Geeraert, Newman & Baayen 2017; Pham & Baayen 2015).

The current dissertation will investigate both production and comprehension of Estonian. One one hand, the large number of inflected forms in Estonian makes the language seem suitable for a rule driven approach in order to decrease the number of forms one must remember. On the other hand, the observations about complex relations in Estonian inflectional paradigms and previous research from other languages in line with these observations, suggest an alternative to this view. Before introducing the individual studies on Estonian, an overview of previous research on whole-word and paradigmatic effects for morphologically complex words is provided.

1.4 Previous research on whole-word frequency and paradigmatic effects

A large portion of research on morphological has been driven by the search for the psychological reality of morphemes, as also indicated by the architecture of most psycholinguistic models mentioned above. Other research, however, has reported whole-word frequency effects for regular complex words, morphological family size effects for both simple or complex words, and inflectional entropy and relative entropy effects for inflected forms in various languages.

Whole-word frequency effects for complex words have been found in visual word recognition. For instance, Baayen et al. (1997) studied plural noun forms (e.g., *handen* 'hands') in Dutch, and found that the frequency of nominal plural form was predictive of processing costs in a visual lexical decision, even when the stem frequency was controlled for. This result was also replicated in Dutch auditory lexical decision task. Baayen et al. (2003) reported that the frequency of regular plural forms of both nouns and verbs speeded up the response times. Baayen, Wurm & Aycock (2007) conducted a regression analysis of nearly 8,500 complex words in the English Lexicon project (Balota, Cortese, Sergent-Marshall, Spieler & Yap, 2004), and found that these effects were not only restricted to high frequency or plural forms but they also emerge across the frequency span and for all types of complex words. This is in contrast to Alegre & Gordon (1999) who claimed that the frequency limit would need to be 6 per million for the whole-word frequency effects to emerge. Additionally, in reading, Kuperman, Schreuder, Bertram & Baayen (2009) reported that whole-word frequency influences early reading times of Dutch transparent compounds. Similar results have also been found in production, for example, Bien, Levelt & Baayen (2005) established that compound frequency affected production latencies in an associative word production task. Finally, in a more recent study, Caselli, Caselli & Cohen-Goldberg (2016) measured the acoustic durations in conversational English speech, and found that durations were negatively correlated with whole-word frequency of English inflected forms.

The most studied paradigmatic measure is probably morphological family size, which is the count of derived and compound words sharing a word. For instance, *manhood*, *manly*, *fireman* all belong to the same morphological family of *man*. It has been shown for a number of languages that words with more family members are processed faster compared to words with less family members (e.g., for Dutch: Bertram, Baayen & Schreuder 2000; Schreuder & Baayen 1997; English: De Jong, Schreuder & Baayen 2003; Finnish: Moscoso del Prado Martín, Bertram, Häikiö, Schreuder & Baayen 2004 and Hebrew: Moscoso del Prado Martín, Deutsch, Frost, Schreuder, De Jong & Baayen 2005). This effect is usually understood as semantic in nature, words that have more family members receive more activation, as they are a part of a larger connected network than words with fewer family members. The cohesion through the shared root within the network also enhances the activation of a particular word, thereby there is no need for the parts of the word to be explicitly segmented (De Jong, 2002). The evidence for the semantic nature of the morphological family size effect comes from research that shows the effect is driven by semantically transparent or relevant subsets of family members. In a study with large Finnish morphological families, Moscoso del Prado Martín et al. (2004) showed that only directly semantically related family members contribute to the facilitatory effect of morphological family size, whereas less related members cause a ceiling effect. For example, in the family for *työ*, there is a facilitatory effect for a dominant family members like *työläinen* 'worker', whereas there would not be an effect for a less dominant members like *urotyö* 'heroic deed'. In another study with Dutch-English bilinguals, Mulder et al. (2014) found that only semantically relevant subsets of family members facilitated processing, and cognates inhibited the processing. For example, in a lexical decision task, the word *tent*, which has the same meaning in both languages, resulted in slower reaction times.

Whereas derived and compound words form a morphological family, inflected forms of the same word form an inflectional paradigm. Previous studies have found that the frequency distribution of the word forms, as well as the size of the inflectional paradigm, affect lexical processing. Inflectional entropy refers to the average amount of information in an inflectional paradigm. The more uniformly the inflected forms are distributed within a paradigm based on their frequency, the higher the inflectional entropy. For example, in English, inflectional entropy is higher for the word *eye* than for the word *nose*, one encounters both the singular *eye* and the plural *eyes*, but rarely the plural form *noses*.

Inflectional entropy has yielded opposite effects for production and comprehension. In Serbian lexical decision experiments, inflected forms with higher inflectional entropy were recognized faster (Milin, Filipović Durdević & Moscoso del Prado Martín, 2009; Moscoso del Prado Martín, Kostić & Baayen, 2004). In contrast, in a Dutch picture naming task, Tabak, Schreuder & Baayen (2010) found inhibitory effects of inflectional entropy. This has been explained by the fact that in picture naming task, one has to make a choice for a particular form in order to be able to produce it, whereas in lexical decision all forms can facilitate processing. The more similar the frequencies of the forms in the paradigm, the harder it is to select the correct one. At the same time, in a task such as lexical decision, one does not have to make a distinctive choice between individual inflected forms within the paradigm. A related measure to inflectional entropy is relative entropy (Kullback-Leibler divergence) which measures the degree of divergence between the frequency distributions of a certain inflectional paradigm and the inflectional class it belongs to. Studies on Serbian and English have found that words in typical paradigms are processed faster than in atypical paradigms (Baayen et al., 2011; Milin et al., 2009).

Finally, a related paradigmatic measure, inflectional paradigm size, has been less studied compared to inflectional entropy. For most languages inflectional entropy is a more informative measure because inflectional paradigms are rather small, thus, the size of individual paradigms cannot vary much. The inflectional paradigm size is an estimation of the number of different non-zero inflectional variants available for a given paradigm in a representative corpus. For example, for English the maximum number of inflected forms in a single paradigm is only two (e.g., *man* and *men*). in contrast, the number of possible forms in Estonian is much higher as shown in Table 1.1 and 1.2. Inflectional paradigm size has received relatively little attention in research and to our knowledge has only been investigated by one study so far. In an Italian lexical decision experiment with verbs and adjectives, Traficante & Burani (2003) reported inhibitory effects for inflectional family size, such that adjectives with more paradigm members were recognized slower than verbs with less paradigm members. However, this result might reflect the difference in processing adjectives and verbs rather than the inflectional family size effect. For example, as shown by Kauschke & Stenneken (2008) verbs have a disadvantage in processing as opposed to nominal categories.

Importantly, the type of inflectional paradigm size count applied in the current dissertation is different than the one considered before. First, unlike in the previous research, it is concerned with paradigm members within nouns and not between different word categories. Second, we consider only the number of forms Estonian speakers actually use, counted as a number of different forms found for a given paradigm in a representative Estonian corpus. One of the largest paradigms in Estonian is the paradigm for *jalg* 'foot/leg' (Table 1.2) which in principle has 46 members containing multiple parallel forms for a plural case. In general, the number of inflected forms used with *jalg* is actually only 36, which has to do with the meaning of the word *jalg*. For example, it is difficult to think of a context where the singular essive case (*jalana* 'as a foot') of *jalg* would make sense, whereas, *jalgadel*, *jalul*, *jalgel* are all frequently used. However, most Estonian paradigms are actually much smaller and used only with a smaller number of cases. Which cases are used, as well as the number of cases a certain word has, are dependent on the semantics of the word. For instance, English mass words such as *furniture* or *jewellery*, lack plural forms. In a corpus study of Finnish, a Finno-Ugric language closely related to Estonian, Karlsson (1986) made a similar observation, he noted that the noun *vesi* 'water' occurs mostly in a series of locative cases, such as the singular inessive *vedessä* 'water' or the singular adessive *vedellä* 'water'. This has to due to with the fact that in the context of *water*, something is often *in water* or *on/with water*. As other words are used in more or less similar contexts, the paradigm size in Finnish or Estonian can largely vary. In the current dissertation, we investigate whether the variation in the number of inflectional paradigm members available in a corpus, i.e,. a word's actual paradigm size, influences Estonian lexical processing. Before turning to an overview of individual studies, we will introduced our three main research questions, methodology and corpus measures.

1.5 Methods and Individual Studies

The current dissertation investigates the production and comprehension of Estonian inflected forms by addressing the following research questions.

1. What kind of information is used by Estonian native speakers when they read and produce inflected forms?
2. When during the time-course does this information become available?
3. Do these lexical effects manifest themselves the same way for all participants?

1. What kind of information is used by Estonian native speakers when they read and produce inflected forms?

Estonian is a morphologically complex Finno-Ugric language with millions of possible complex word forms, which is particularly known for its rich inflectional system with 14 cases in singular and plural. A rule-driven morpheme based account seems beneficial for languages like Estonian, as it would reduce the number of forms one must memorize. Indeed this has been the traditional position in both psycholinguistic (Hankamer, 1989; Niemi et al., 1994; Yang, 2016) and computational (Hankamer, 1989; Kaalep, 1997; Karlsson & Koskenniemi, 1985; Sproat, 1992) approaches to rich inflectional systems, but also inflectional morphology in general (see e.g., Halle & Marantz 1993; Pinker 1999; Taft 1994). However, as discussed above, recent research challenges this view and shows that various whole complex word level predictors, such as whole-word frequency (Baayen et al., 1997; Balling & Baayen, 2008; Caselli et al., 2016; Kuperman et al., 2009), inflectional entropy (Milin et al., 2009; Moscoso del Prado Martín et al., 2004; Tabak et al., 2010) and morphological family size (De Jong, Feldman, Schreuder, Pastizzo & Baayen, 2002; Moscoso del Prado Martín et al., 2004; Schreuder & Baayen, 1997) also co-determine lexical processing costs. Importantly, however, these findings come from languages that are morphologically less productive than Estonian. This raises the question of whether such effects will also emerge and generalize to morphologically very rich languages, such as Estonian. In addition to the more well-established whole-word frequency, morphological family size and inflectional entropy effects, we also investigate whether inflectional paradigm size, as measured by the number of inflected forms in the paradigm in actual use, co-determines processing costs in Estonian.

2. When during the time-course does this information become available?

Our second research question is concerned with the time-course of Estonian inflectional processing, that is, when does certain information about complex words become relevant. According to Fruchter & Marantz (2015) and Taft (2004), the initial stages of processing are driven by morpheme-based processes, processes, while whole word effects arise at later stages of processing when morphemes are combined in order to compute the meaning of the word. However, in a series of priming experiments, Feldman and colleagues were able to show that the time-course of processing is early on depend on the semantics of the complex word (Feldman, 2000; Feldman et al., 2009; Feldman & Soltano, 1999). Further, Kuperman et al. (2009) conducted an eye-tracking experiment with isolated Dutch compounds and observed that the frequency of the whole compound facilitated processing as early as during the first fixation. Finally, an early temporal locus of whole-word frequency effect was recently supported by Schmidtke et al. (2017). Using distributional survival analysis of the lexical decision times, they observed that the whole-word frequency emerges as a significant predictor of lexical decision latencies much earlier in time than constituent frequency. The quantile regression analysis in Study 2 and the eye-movement analysis in Study 3 explore the time-course of lexical processing in Estonian.

3. Do these lexical effects manifest themselves the same way for all participants?

Our final research question is concerned with individual differences in morphological processing. More precisely, we are interested in whether and how actual life-long exposure to the complexity of the Estonian system predicts speaker's comprehension and production of complex forms. The research on individual differences in morphological processing has been relatively sparse, and studies investigating this have only recently started to emerge (Andrews & Lo, 2013; Falkauskas & Kuperman, 2015; Kuperman & Dyke, 2011; Schmidtke, Van Dyke & Kuperman, 2017).

In the current work, we explore individual differences from two perspectives. First, we examine whether younger and older Estonian native speakers process complex words differently by determining whether participants' age is correlated with their response times and eye movements patterns, and additionally, if age interacts with lexical properties when reflecting participants' responses. Older participants have inherently had more exposure to the language (Ramscar, Hendrix, Shaoul, Milin & Baayen, 2014; Ramscar, Sun, Hendrix & Baayen, 2017), which we expect to affect their language processing. A range of lexical decision, naming and reading studies, which require speeded reaction, have found that older participants are slower and make more mistakes in production and reading tasks which require speeded reaction, whereas older participants' semantic abilities do not seem to decline and can even be advantagious compared to younger participants (see e.g., Burke & Shafto 2008). In addition, more precise differences between older and younger participant have been studied. In a word naming task, Spieler & Balota (2000) showed that compared to younger participants, older participants were more facilitated by word frequency, and less facilitated by orthographic neighbourhood and length compared to younger participants. This suggests that older people might have a more unitized representation of words. Larger frequency effects for older people have also been reported using other paradigms, such as reading (Rayner, Reichle, Stroud, Williams & Pollatsek, 2006) and lexical decision (Allen, Madden & Crozier, 1991).

Participants in all our studies were from a wide age range, and this factor was included as a covariate in the first three studies. In Study 4, individual differences will be investigated from a different perspective, by comparing processing effects between participants and participant groups using pupillometry.

1.5.1 Experimental methodology

The present dissertation takes a multi-methodological approach to studying the processing of Estonian case-inflected nouns. The following methods were used: 1) visual lexical decision (response time and error analyses), 2) word naming (production latencies and acoustic duration analyses), 3) semantic categorization (response time and error analyses), and 4) eye tracking (eye fixation analysis and pupillometry). We used both traditional behavioural measures, such as naming latencies and durations, and lexical decision times and accuracy, as well as eye tracking measures, such as fixation times and pupil dilation as an indicator of online processing. Unlike decision-based measures, eye tracking provides a more fine-grained insight (e.g., first fixation duration, total fixation duration) into the time-course of inflectional processing.

We will use both more traditional gaze-based measures, as well as pupillometry. Previous research has shown that changes in pupil diameter reflect changes in emotional state and cognitive load (Ahern & Beatty, 1979; Hess & Polt, 1960; Kahnemann & Beatty, 1966). These studies indicate that pupil dilates not only in response to changes in lighting conditions (Young & Biersdorf, 1954), but also in response to changes in mental effort. Pupil dilation has also been used in linguistics in order to gauge to the complexity of the task (Hyönä, Tommola & Alaja, 1995), sentence intelligibility (Zekveld, Kramer & Festen, 2010), and sentence complexity (Ben-Nun, 1986; Engelhardt, Ferreira & Patsenko, 2010a; Just & Carpenter, 1993; Schluroff, Zimmermann, Freeman, Hofmeister, Lorscheid & Weber, 1986).

A few studies have also used pupillometry in lexical processing research (Geller, Still & Morris, 2016; Kuchinke, Võ, Hofmann & Jacobs, 2007; Papesh & Goldinger, 2012). For instance, Kuchinke et al. (2007) showed that pupil response is modulated by frequency in an English lexical decision task and that lower frequency words receive a stronger pupil response. Papesh & Goldinger (2012) also showed this in a delayed word naming task. Geller et al. (2016) studied the effect of orthographic similarity with a priming task and found that pupil response was stronger when prime-target pairs were orthographic similar.

However, all these studies have looked at only the peak pupil dilation as opposed to the time-course of pupil dilation across the trial. Furthermore, they assumed the same pupil response pattern to the task from all participants, hence discarding individual differences. We explored pupillometry in new ways compared to previous studies. First, we studied the time-course of pupil dilation across the trials as opposed to just measuring the pupil peak. Second, we did not presuppose similar effects in pupil size for all participants, and in addition to overall pupil analysis, we also conducted individual participant and participant group level analyses.

1.5.2 Statistical methods

Throughout this dissertation, the data was analysed using Generalized Additive Mixed Models (GAMM, Wood 2006, R-package *mgcv*, see Baayen, Vasishth, Bates & Kliegl 2017). We opted for GAMM analysis, as opposed to other regression techniques, such as Linear Mixed Effect Models (Baayen, 2008; Baayen, Davidson & Bates, 2008), because GAMM-s do not assume linearity between the predictor and response variable. In recent years, GAMM-s have gained popularity in analysing various type of linguistic data such as EEG data (Tremblay & Baayen, 2010), dialectometry (Wieling, Nerbonne & Baayen, 2011), reaction times (Feldman, Milin, Cho, del Prado Martín & OConnor, 2015; Pham & Baayen, 2015), speech production (Kösling, Kunter, Baayen & Plag, 2013; Tomaschek, Wieling, Arnold & Baayen, 2013) and spoken word perception (Porretta, Tucker & Järvikivi, 2016).

The output of GAMM consists of two parts: a parametric part and a non-parametric part. The parametric part provides information about the intercept and the linear coefficient (slope), which is identical to that of standard linear models. The intercept is the expected mean value of the response variable, when predictor variables are zero. The non-parametric part covers a range of very different effects:

non-linear phenomena, interactions between predictors and random effects (intercept adjustments, random smooths and random slopes). The non-parametric part does not provide coefficients for the effects, therefore the direction and the size of the effect can be best estimated through visualization.

Additionally, in chapters 3 and 4, we complemented the GAMM analysis with quantile regression (R-package *qgam* by Fasiolo, Y., Nedellec & Wood 2016). Using this statistical method is relatively new in psycholinguistic analysis, however, it provides several potential benefits. First, whereas GAMM provides estimates for the mean value of response variable only, quantile regression also provides estimates for low and high quantiles of the response variable. Thus, we can investigate how very short and long response times are affected by measures such as frequency, length or paradigm size. Second, compared to GAMM-s, quantile regression is a distribution free regression technique, so it does not require the response variable to be normally distributed. Lastly, quantile regression is not affected by autocorrelation in the model residuals, which will be particularly beneficial in dealing with the complex and highly variable nature of pupil data.

1.5.3 Estonian corpus measures

Our main lexical and lexical-distributional variables used across studies are summarized below. Most corpus measures were calculated on the basis of the 15-million token Balanced Corpus of Estonian [4], which is the best morphologically tagged and publicly available Estonian corpus. The corpus contains newspaper texts, fiction texts and scientific texts in equal parts (5 million words each). Morphological family size was calculated based on the online version of the Estonian Word Families dictionary (Vare, 2012)[5]. Estonian Word Families dictionary contains information about the structure and relations of 9,000 Estonian word families.

Orthographic neighbourhood density. Orthographic neighbourhood density counts words which differ by one grapheme from the target word. This was calculated using the function *levenshtein.distance* from the R-package *vwr* (Keuleers,

[4] http://www.cl.ut.ee/korpused/grammatikakorpus/ (15.04.2017)

[5] `http://www.eki.ee/dict/sp/`

2013). For instance, the orthographic neighbourhood density for the elative plural *jalust* 'from feet' is 6, and includes words like the partitive plural *alust* 'base', the elative singular*jahust* 'from/of flour', the elatice singular *janust* 'from thirst', the elative singular *salust* 'from/of the bosk', the elative singular*talust* 'from/of the cottage', and the elative singular *valust* 'from/of pain'. We excluded orthographic neighbours within the same inflectional paradigm from this count (e.g., the elative singular *jalast* 'from the foot').

Orthographic word length. Orthographic word length varied between 2 and 35 characters (mean 11.51; sd 3.88 characters) in the corpus. One of the many two-character words in the corpus was the monomorphemic word, *au* 'honour'; the longest polymorphemic word with many derivational and inflectional affixess was *sisekergejõustikumeistrivõistlustel* 'at the indoor championships in athletics'. Estonian has a shallow orthography, hence the number of graphemes closely corresponds to the number of phonemes.

Whole-word frequency. Whole-word frequency accounts for the total number of tokens in a corpus for a given word form. For example, the nominative plural form *jalad* 'feet' occurs 1067 times, the elative singular form *jalast* 'from a foot' 149 times, and the adessive plural form *jalgadel* 'on the feet' 80 times in the Balanced Corpus of Estonian.

Lemma frequency. Lemma frequency is the cumulative frequency of the whole inflectional paradigm. For example, the summed frequency of the lemma *jalg* 'foot' is 5193, and includes forms like the nominative singular*jalg*, the genitive singular *jala*, the partitive singular *jalga* 'partitive singular'.

Inflectional paradigm size. This measure counts the number of inflectional variants in use for a certain lemma, calculated from a representative corpus.. For example, the paradigm size for jalg 'foot; leg' is 36 in the Balanced Corpus of Estonian. All syncretic and parallel forms were included separately in the inflectional paradigm size measure. Forms with clitics were excluded from the count (e.g., *jalgadelgi* 'even on the feet').

Inflectional entropy. Inflectional entropy measures the average amount of information in an inflectional paradigm. The higher the entropy, the more uniformly the inflected forms are distributed with respect to frequency of occurrence, and as a result, the higher the uncertainty (or, the lower the information value). Inflectional entropy (*HI*) was calculated using a formula adapted from Moscoso del Prado Martín et al. (2004):

$$\mathrm{HI}(w) = -\sum_{i=1}^{n} p(w_i)\log_2 p(w_i) = -\sum_{i=1}^{n} \frac{f(w_i)}{f(w)}\log_2\frac{f(w_i)}{f(w)}$$

where *w* is the lemma, *n* the number of inflected variants, *f(w)* lemma frequency, the summed frequency of variants, *f(w_i)* is the frequency of an particular invariant, and *p(w_i)* the probability within the paradigm probability the given variant. For instance, inflectional entropy for the *kriim* 'scratch' paradigm is relatively high 3.53, whereas entropy for *aru* 'mind' is low 1.31, whereas both have 10 paradigm members.

Morphological family size. Morphological family size is the number of compound and derived words that share a constituent. For instance, the morphological family size for *jalg* is 363 and it includes forms like *jalgpall* 'football', *jalgsi* 'afoot', *lampjalg* 'flatfoot'. Estonian morphological family size counts were retrieved from the online version of the Estonian Word Families dictionary.

1.5.4 Specific studies

The research in the current dissertation will be presented as separate, journal-style papers. The first study investigates the comprehension of Estonian case-inflected nouns using lexical decision and semantic categorization tasks. The second study addresses the production of Estonian with two word naming tasks. The third study explores the time-course of Estonian inflectional processing during word naming using both gaze data and pupillometry. Finally, the fourth study concentrates on individual differences using pupillometry. A short summary of each of these study is given below.

Study 1

Study 1 investigated the comprehension of Estonian case-inflected nouns with a visual lexical decision task (3,000 items; 24 participants, age 24-67 years) and a semantic categorization task (200 items; 26 participants, age 21-67 years). A number of variables, such as lemma frequency, whole-word frequency, inflectional paradigm size, inflectional entropy and morphological family size emerged as predictors of Estonian lexical processing costs. Experiment 1 documented that lexical decision response times decreased and accuracy scores increased with growing whole-word frequency, inflectional paradigm size (the number of in a corpus attested inflectional variants), and with decreasing orthographic length. Lemma frequency also emerged as a significant predictor, but including lemma frequency in the model resulted in a worse fit to the data than including whole-word frequency. In addition, we found an interesting trade-off, showing that both response times and accuracy increase with age. Experiment 2 replicated the inflectional paradigm size effect in Experiment 1 with an animate-inanimate categorization task. The particular focus was on inflectional paradigm size, as we predicted that if inflectional paradigm size effect is semantic in nature, it should also persist in a semantic categorization task. We found that response times and accuracy were indeed modulated by inflectional paradigm size. Inflected forms with higher paradigm size were categorized faster and more accurately. In line with the previous experiment, older participants were slower but more accurate at the task. Additionally, our results showed that animate nouns were categorized faster and more accurately than inanimate nouns and that shorter words were recognized faster and more accurately than longer words. The whole-word frequency effect found in Experiment 1 is surprising from the perspective of accounts relying on obligatory morphemic decomposition. However, it fits well with the evidence for frequency effects for English multi-word sequences. Finally, the inflectional paradigm size effect is line with the literature on a similar measure, morphological family size.

Study 2

Study 2 investigated the processing of Estonian inflection with a word naming task. In Experiment 1 (200 items; 26 participants, age 21-67 years), we observed that words with larger morphological families elicited shorter production latencies than words with smaller families, and that forms with higher whole-word frequencies elicited shorter acoustic durations lower whole-word frequency. Experiment 2, for which we increased the statistical power by using 2,800 words (33 participants, age 22-69 years) revealed that higher whole-word frequency, inflectional paradigm size, as well as morphological family size all reduced both production latencies and acoustic durations. Additionally, a relatively new statistical method, quantile regression, showed that whole-word frequency, inflectional paradigm size and morphological family size were predictive of both short and long production latencies. The frequency effect was most predictive for short production latencies, reflecting early processing and paradigmatic effects, whereas inflectional paradigm and morphological family size were most predictive for longer production latencies, reflecting later processing. Hence, Study 2 is in line with findings from previous study, replicating the inflectional paradigm size effect also for production. Additionally, the current study established the morphological family size effect in the acoustic durations of inflected forms, as the first to our knowledge.

Study 3

Study 3 investigated the time-course of lexical processing using eye movement and pupil size measurement during a word naming task (Study 2: Experiment 2). The aim of this study was to investigate whether eye tracking data converges with behavioural data and confirms the earlier locus of whole-word frequency effect and later emergence of paradigmatic effects. As an additional aspect of this study, we explored the possibilities of using pupillometry in studying morphological processing. The results of the study revealed that first fixations were not affected by any other lexical property besides orthographic length. Lexical effects emerged slightly later during the second fixation duration, and this measure was sensitive to whole-word frequency. Further, the total fixation analysis revealed an effect of inflectional paradigm size. Thus, eye fixation analyses were in line with the findings from the two previous behavioural studies in the present dissertation, indicating that whole-word frequency effect might arise earlier and paradigmatic effects later in time. Furthermore, analyses using pupil dilation data supported the evidence from gaze data, but also suggested that there are individual differences in the processing patterns of individual speakers. For instance, orthographic length and participant age effect interacted. Younger participants had an increased pupil dilation compared to older participants when reading long words, which might be a reflection of younger participants having less experience with language than older participants (Ramscar et al., 2014, 2017).

Study 4

Study 4 concentrated on individual differences in pupil dilation during word naming. Study 3 established that pupil size reflects general trends such as the facilitatory effect of inflectional paradigm size and whole-word frequency, but also indicated that there were large individual differences in pupil dilation patterns. Thus, we wanted to further investigate individual differences. First, we observed that although all participants performed a same task, word naming, their pupil dilation patterns over time largely varied. Second, these individual pupil patterns were feed into a hierarchical clustering algorithm to see whether different participant groups can be established. The modelling results showed that three main participant groups exists, and that they all differ with respect to their sensitivity to word's frequency. Group 1 showed both early and late frequency effects, group 2 showed only weak frequency effects overall and finally, group 3 showed late frequency effects, but the effects were much weaker for this group. Interestingly, these groups cannot be reduced to age, thus, further research is needed to establish what triggers differences. In summary, looking at individual pupil curves as opposed to averaging across all participants provides interesting insights into individual differences in participants' pupil patterns.

Chapter 2

Whole-word frequency and inflectional paradigm size facilitate Estonian case-inflected noun processing

Abstract

Estonian is a morphologically rich Finno-Ugric language with nominal paradigms that have at least 28 different inflected forms, but sometimes more than 40. For languages with rich inflection it has been argued that whole-word frequency, as a diagnostic of whole-word representations, should not be predictive for lexical processing. We report a lexical decision experiment, showing that response latencies decrease both with frequency of the inflected form and its inflectional paradigm size. Inflectional paradigm size was also predictive of semantic categorization, indicating it is a semantic effect, similar to the morphological family size effect. These findings fit well with the evidence for frequency effects of word n-grams in languages with little inflectional morphology, such as English. Apparently, the amount of information on word use in the mental lexicon is substantially larger than was previously thought.

2.1 Introduction

Estonian is a Finno-Ugric language with remarkably productive morphology. A 15-million word corpus of Estonian[1] contains no less than 790,957 different words, a number similar to the total number of different words (794,771) in a 100-million word corpus of British English (Leech, Rayson & Wilson, 2014). This raises the question of how a speaker of Estonian can understand such a large number of different forms, especially considering the probability of encountering an out-of-vocabulary word, i.e., a word the speaker has not yet seen or heard, is no less than 0.64.

However, the problem might not be as severe as it may seem, as out-of-vocabulary words are typically morphologically complex. In fact, roughly 95% of the forms in our corpus have morphological structure, including derived words (e.g., *töötaja* 'worker') and compounds (e.g., *käsitöö* 'handwork'). Derived words and compounds built on the same stem cluster into morphological families, while inflected forms (e.g., *tööd* 'works', *töös* 'at work', *tööga* 'with the work') cluster into inflectional paradigms. Inflectional paradigms typically come with a few inflected variants, the so-called principal parts, from which all other forms in the paradigm can be predicted (Blevins, 2006). Accordingly, the number of forms that one must know by heart is much smaller than the number of forms that one can understand or produce, given these basic forms and the rules of the language.

Several studies have argued that for morphologically rich languages, such as Finnish and Turkish that storing all word forms in a mental dictionary would exceed the storage capacity of the brain (Hankamer, 1989; Niemi, Laine & Tuominen, 1994; Yang, 2016). Additionally, mental dictionaries with only stored forms would not be able to deal with the large numbers of out-of-vocabulary words. Therefore, algorithms must be available for interpreting and producing novel complex words, both in natural language processing systems and in the human cognitive system (Hankamer, 1989; Kaalep, 1997; Karlsson & Koskenniemi, 1985; Sproat, 1992).

Although morphological decomposition has also been argued to play a funda-

[1] http://www.cl.ut.ee/korpused/grammatikakorpus/ (15.01.2017)

mental role even in languages with simple morphologies such as English (Fruchter & Marantz, 2015; Rastle et al., 2004; Taft, 1994; Taft & Forster, 1975), it is specifically for languages with rich morphology such as Estonian in order to minimize storage and maximize rule-driven computation (Pinker, 1999) seems especially attractive.

Even though Estonian appears to be a prime candidate for a language dominated by rule-driven processing, recent findings place Estonian morphology in a different light. For languages as diverse as English and Mandarin, experimental evidence is accumulating that the frequency of occurrence of sequences of multiple words (e.g., *the president of the*) predicts a range of measures of lexical processing when other predictors, such as word frequency and length, are statistically controlled (Arnon & Snider, 2010; Janssen & Barber, 2012; Sun, 2016; Tremblay, Derwing, Libben & Westbury, 2011; Tremblay & Tucker, 2011). These frequency effects have not only been found in studies with adults, but also in studies with children (Ambridge, Kidd, Rowland & Theakston, 2015; Bannard & Matthews, 2008; Kidd, 2012) and second language learners (Siyanova-Chanturia, Conklin & Van Heuven, 2011; Sonbul, 2015; Wolter & Gyllstad, 2013). Importantly, sequences of words in English, such as *into to the house*, translate into Estonian with a single inflected form, such as *majasse*. In the light of these frequency effects for English, we predict a similar frequency effect for functional equivalence in Estonian.

Given the frequency effects for word sequences, it is unsurprising that whole-word frequency effect in the processing of regular complex words are also attested (Dutch: Baayen, Dijkstra & Schreuder 1997; Baayen, McQueen, Dijkstra & Schreuder 2003; Kuperman, Schreuder, Bertram & Baayen 2009; English: Baayen, Kuperman & Bertram 2010; Baayen, Wurm & Aycock 2007; Vietnamese Pham & Baayen 2015; and Danish: Balling & Baayen 2012). For Finnish, a Finno-Ugric language closely related to Estonian, whole-word frequency effects have been found for derived words, however, not for most inflected forms (Bertram, Laine, Baayen, Schreuder & Hyönä, 1999; Laine, Vainio & Hyönä, 1999; Niemi, Laine & Tuominen, 1994; Soveri, Lehtonen & Laine, 2007; Vannest, Bertram, Järvikivi & Niemi, 2002). One reason may be that inflection rather serves syntactic functions, such as grammatical role, number marking and agreement, whereas derivation and compounding typically serve the formation of new names for things and events. However, a problem with previous studies on inflected forms in Finnish is the small number of subjects and items, as well as the concomitant lack of power (Westfall, Kenny & Judd, 2014). Thus, the first goal of the present study was to re-examine whole-word frequency effects for inflected words in Estonian using a large regression design with thousands of items.

The consequences for lexical processing of the size of a word's morphological family e.g., *worker*, *workforce*, *handwork*) have been studied extensively (Bertram et al., 2000; De Jong, Schreuder & Baayen, 2000; Moscoso del Prado Martín et al., 2004; Schreuder & Baayen, 1997). Words with larger families are processed faster, which has been explained in two ways. Within the framework of interactive activation (De Jong et al., 2003), words from larger families receive more activation from their family members, resulting in a critical threshold activation level being reached earlier in time. According to learning models (e.g., Baayen et al. 2011), as long as complex forms share some element of meaning, that element will be strengthened for all the family members each time it is encountered, allowing for a faster reaction time for words with larger families. The morphological family size effect is generally understood as a semantic effect, as it appears to be driven primarily by semantically transparent family members or semantically relevant subsets of family members (Moscoso del Prado Martín et al., 2004; Mulder et al., 2014). As semantic transparency is greater for inflection as compared to derivation and compounding, an effect of inflectional paradigm size should be detectable for languages with large inflectional paradigms.

Only a few studies have looked at the role of inflectional paradigms in lexical processing. Moscoso del Prado Martín et al. (2004) studied the processing consequences of inflectional paradigms in English and Dutch, using summary measures characterizing the probability distribution of inflectional variants. Specifically, inflectional entropy and the Kulback-Leibler divergence have been found to predict the consequences of inflectional paradigmatic relations in the lexical decision task (Milin et al. 2009, see also Baayen et al. 2011 for prepositional entropy effects for English).

The size of Estonian nominal paradigms offers further opportunities for investigating the consequences of paradigm complexity. Estonian nominal paradigms have 14 cases in both singular and plural, but may have several additional parallel forms. However, in practice most words are actually not used in all their cases and numbers. For example, for *jalg* 'foot, leg' out 46 grammatically possible forms only 36 inflected forms are present in the Balanced Corpus of Estonian (Table 2.1).

Table 2.1: Inflectional paradigm of jalg 'foot, leg' with 46 paradigm members. The 36 paradigm members present in the corpus are marked in bold.

Case	Singular	Plural	English translation
Nominative	**jalg**	**jalad**	foot (subject)
Genitive	**jala**	**jalgade**, **jalge**	of a foot; foot (as a total object)
Partitive	**jalga**	**jalgasid**, **jalgu**	foot (as a partial object)
Illative-1	**jalga**	-	into a foot
Illative-2	jalasse	**jalgadesse**, jalusse, jalgesse	into a foot
Inessive	**jalas**	**jalgades**, **jalus**, **jalges**	in a foot
Elative	**jalast**	**jalgadest**, **jalust**, **jalgest**	from a foot
Allative	**jalale**	**jalgadele**, **jalule**, **jalgele**	onto a foot
Adessive	**jalal**	**jalgadel**, **jalul**, **jalgel**	on a foot
Ablative	**jalalt**	**jalgadelt**, jalult, **jalgelt**	from a foot
Translative	**jalaks**	**jalgadeks**, jaluks, jalgeks	[to turn] into a foot
Terminative	jalani	**jalgadeni**, jalgeni	up to a foot
Essive	jalana	jalgadena	as a foot
Abessive	**jalata**	**jalgadeta**	without a foot
Comitative	**jalaga**	**jalgadega**	with a foot

In a Finnish corpus study, Karlsson (1986) made a similar observation, pointing out that although in theory a word can appear in any of the inflected forms defined by grammar, only a subset of the possible forms actually occurs. Figure 1.1 illustrates this point for Estonian. Most paradigms only occur with one member, but there are also paradigms with up to 38 members.

Karlsson (1986) argues that the forms that end up being attested (i.e. forms that actually get used), depends on the meaning of the word. For example, *kesä* 'summer' has mostly a temporal meaning, and therefore the most frequent inflected form is the adessive *kesällä* 'in the summer', whereas other forms are used less frequently or not at all. Likewise, although in theory the number of paradigm

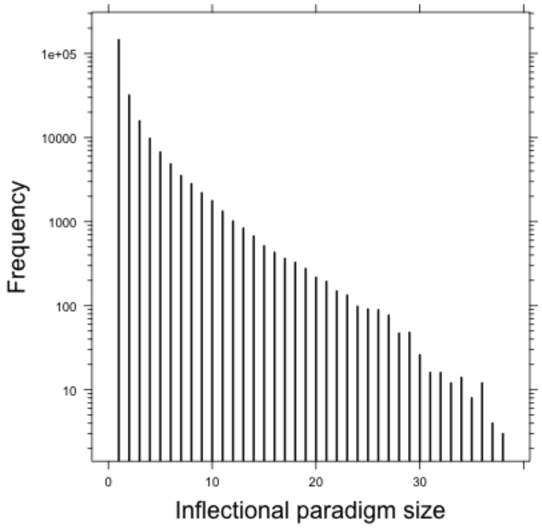

Figure 2.1: Histogram of 231,891 noun paradigms in the Balanced Corpus of Estonian, 75% of which are paradigms of compound. The x-axis represents inflectional paradigm size, the y-axis shows the frequency of a particular paradigm size.

members for Estonian nouns can be quite high, in practice not all possible inflected forms are semantically felicitous. This observation brings us to our second research goal, namely whether the actual size of Estonian nominal paradigms has consequences for lexical processing. Specifically, does a large inflectional paradigm size facilitate processing, just as large morphological families do? Experiment 1 makes use of a visual lexical decision task to address these issues.

2.2 Experiment 1: Lexical decision task

2.2.1 Participants

Twenty-four native speakers of Estonian (14 females; 24-67 years, mean 43.69) with normal or corrected-to-normal vision completed the experiment. They received 10 euros for their participation.

2.2.2 Materials

1,000 nouns were retrieved from the Balanced Corpus of Estonian. For each noun, three different case-inflected forms (either nominative, genitive, partitive; inessive, allative, ablative; translative, essive, comitative singular, and nominative or two variants of plural illative case) were selected. The whole-word frequency distribution of stimulus set resembled that of the corpus, and ranged between 1 and 3,402 occurrences per million (median 4). Length in letters varied between 3 and 21 characters (mean 7 characters). Inflectional paradigm size ranged between 2 and 36 (median 19).

From the list of 3,000 words, twelve experimental lists were created, each contained 250 words in randomized order. Each list was enlarged with 250 inflected nonwords, created by changing one or two letters of existing Estonian inflected forms while respecting Estonian orthophonotactics. Each subject was presented with 500 items. Across the experiment, each inflected form received two responses, and each lemma six responses.

2.2.3 Procedure

We used a standard visual lexical decision task. Stimuli were presented on a 15-inch Dell laptop in 18 point Courier New Bold font on a white background, using E-Prime 2.0 experimental software (Psychology Software Tools). Each trial started with the visual presentation of a blank screen for 1,000 ms, after which a fixation point appeared in the centre of the screen for 750 ms, followed by the target stimulus in the same position for a maximum of 2,500 ms. After responding, or after time-out, the stimulus disappeared, and a new trial was triggered. The experiment was divided into five blocks with 100 trials each. The first experimental block was preceded by 20 practice trials. A break followed each block and lasted until the participant was ready to continue. One experimental session lasted approximately 45 minutes.

2.2.4 Analysis and Results

Nonword trials and trials with incorrect responses were excluded from reaction time analysis, additionally trials with response times less than 50 ms were removed prior to analysis from both reaction time and accuracy analyses. We analysed the data with the generalized additive mixed model (GAMM, Baayen, Vasishth, Bates & Kliegl 2017; Wood 2006) using the *R*-package *mgcv* (version 1.8-12), with inversed-transformed reaction time (-1000/RT) as response and as predictors inflectional paradigm size (inflectional variants of a paradigm attested in the corpus), whole-word frequency, lemma frequency (the summed frequency of a noun's inflectional variants), word length and participant's age in years. To avoid outlier effects, frequency was log-transformed. For visualization, we made use of the *itsadug* package (version 2.2, van Rij, Baayen, Wieling & van Rijn, 2016).

Reaction times increased linearly with age ($\hat{\beta} = 0.0077, t(5130) = 3.2921, p = 0.001$). Reaction times decreased nonlinearly for increasing inflectional paradigm size ($F = 6.558, \mathit{edf} = 1.6554, p = 0.0013$). Whole-word frequency and word length entered into a non-linear interaction that was modelled with a tensor product smooth ($F = 46.6386, \mathit{edf} = 4.7394, p < 0.0001$, see Baayen et al. 2017 for an introduction to the generalized additive model). As expected, reaction times decreased with frequency and increased with length. The interaction indicated that the frequency effect was slightly larger for the shortest words.

Whole-word frequency was correlated with lemma frequency ($r = 0.43$). Replacing whole-word frequency by lemma frequency, however, resulted in a model with a substantially worse fit (increase in AIC 147.27), and adding lemma frequency to the model with whole-word frequency revealed lemma frequency not to be significant when whole-word frequency was present as predictor.

The model included by-subject factor smooths for trial ($F = 778.5655, \mathit{edf} = 136.1335, p < 0.001$), as well as by-subject random slopes for length ($F = 11.8300$, $\mathit{edf} = 20.2773, p < 0.0001$), frequency ($F = 2.6188, \mathit{edf} = 13.7509, p < 0.001$), and paradigm size ($F = 1.0644, \mathit{edf} = 10.0954, p < 0.0063$). The model also included by-item random intercepts ($F = 0.2316, \mathit{edf} = 473.2887, p < 0.0001$). Further random effects and interactions did not reach significance.

A logistic GAMM fitted to the accuracy data supported the conclusions drawn from the analysis of reaction times. The probability of an error decreased with inflectional paradigm size ($\chi^2 = 17.3905, \mathit{edf} = 2.7258, p = 0.0010$) and with age ($\hat{\beta} = 0.0206, z(5945) = 2.8442, p = 0.0045$). Whole-word frequency and length showed a strongly non-linear interaction (see Figure 2.2, $\chi^2 = 46.6386, \mathit{edf} = 4.7394, p <$

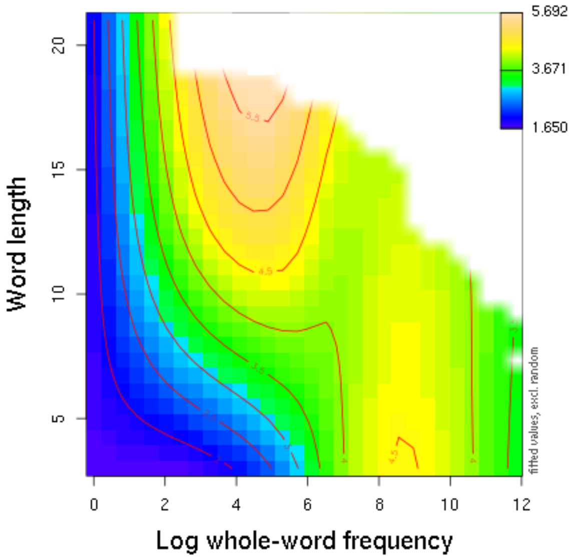

Figure 2.2: Tensor product smooth for the interaction of word length by word frequency. Colour coding is used to represent model predictions, with darker shades of yellow indicating higher log-odds for correct responses, and darker shades of blue representing lower log-odds ratios.

0.0001): accuracy increased with whole-word frequency and word length, but was especially high for long words of intermediate frequency. The random effect structure was the same as for the reaction times (all $p < 0.05$).

Given that the morphological family size effect appears to be semantic in nature and given that most cases in Estonian also express semantic meanings, such *jalal* 'on the foot' or *jalaga* 'with the foot', it seems likely that an inflectional paradigm size effect would also be a semantic effect. Experiment 2 addresses this issue by means of a semantic categorization task. We asked participants to decide whether words referred to animate or inanimate objects, as speakers tend to have good natural intuition regarding what is animate. If an inflectional paradigm size effect indeed exists, and if it is semantic in nature, then it should persist in semantic categorization. By contrast, the whole-word frequency effect may well be absent in this task, given the results reported by Balota & Chumbley (1985).

2.3 Experiment 2: Semantic categorization task

2.3.1 Participants

26 native speakers of Estonian (18 female; age 21-67 years, mean 38.66) with normal or corrected-to-normal vision were recruited. They received 10 euros for their participation.

2.3.2 Materials

200 case-inflected nouns were selected from the Balanced Corpus of Estonian, 100 animate and 100 inanimate nouns. The stimuli were selected in such a way that the correlation between the inflectional paradigm size and whole-word frequency was low ($r = 0.3$). However, the correlation between inflectional paradigm size and lemma frequency remained high ($r = 0.8$). Whole-word frequency ranged between 1 and 213 per million (median 5), length in letters varied between 4 and 15 characters (mean 8 characters), and inflectional paradigm size ranged between 2 and 36 (median 22). The same 200 items were presented to all participants in a randomized order.

2.3.3 Procedure

Participants classified stimuli as animate or inanimate by pressing the relevant computer key. Stimuli were presented on a 17-inch Dell computer screen in 26 point Courier New Bold font on a light grey background, using ExperimentBuilder software (SR Research Ltd). Each trial started with the visual presentation of a blank screen for 1,000 ms, followed by a fixation cross for 500 ms, after which the word appeared at the center of the screen for 2,500 ms or until a decision was made. The experiment started with six practice trials, followed by 200 experimental trials. The experiment took approximately 20 minutes.

2.3.4 Analysis and Results

A GAMM with log-transformed reaction time as response variable revealed that inanimate nouns were responded to more slowly than animate nouns ($\hat{\beta} = 0.0528$,

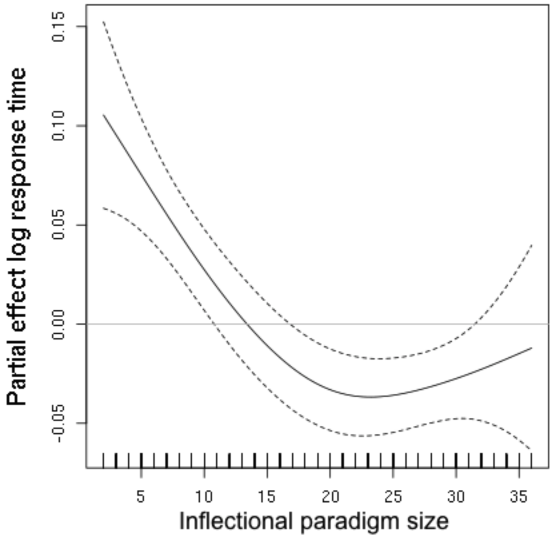

Figure 2.3: The partial effect for inflectional paradigm size in semantic categorization task. Increasing inflectional paradigm size decreases response times. The effect disappears for larger paradigm sizes.

$t(4527) = 2.4057, p = 0.0162$). Reaction times increased linearly with age ($\hat{\beta} = 0.0074, t(4527) = 3.3253, p = 0.0009$), and nonlinearly with word length ($F = 5,796, \mathit{edf} = 3.1301, p = 0.0007$), and decreased with inflectional paradigm size ($F = 10.9108, \mathit{edf} = 2.6432, p < 0.0001$). As shown in Figure 2.3, the effect of paradigm size is restricted to the smaller paradigm sizes. Effects of word frequency or lemma frequency were not significant.

The model included by-subject factor smooths for trial ($F = 215.8889, \mathit{edf} = 120.8497, p < 0.0001$), as well as by-subject random slopes for length ($F = 3.2682$, $\mathit{edf} = 17.6420, p < 0.0001$), and condition ($F = 2.9903, \mathit{edf} = 21.5936, p < 0.0001$). The model also included by-item random intercepts ($F = 3.3404, \mathit{edf} = 146.6001, p < 0.0001$). Further random effects and interactions did not reach significance.

A logistic GAMM fitted to the accuracy data did not support the effects of order, age, and length. However, a larger paradigm size predicted fewer errors ($\hat{\beta} = 0.0313, z(5058) = 2.1712, p = 0.0299$), and inanimate nouns elicited less errors ($\hat{\beta} = 0.9850, z(5058) = 3.0472, p = 0.0023$). The random effect structure was the same as for the reaction times (all $p < 0.001$).

2.4 General Discussion

Even though Estonian has highly productive and complex inflectional morphology, making it a prime candidate for decompositional processing, Experiment 1 showed that whole-word frequency was a better predictor than lemma frequency. This dovetails well with our second finding that inflectional paradigm size also predicted reaction times in both experiments.

Are these effects due to irregularities in Estonian paradigms? Estonian has not only agglutinative inflected forms, but also fusional forms, e.g., *kalu* 'partitive plural fish', in which lexical meaning, number and partitive are all bundled together in a non-decompositional way. Case-inflected forms are often used as lexicalized adverbs, e.g., *käes* 'there', literally 'in the hand'. Furthermore, many paradigm members have alternative parallel forms, which may express subtly different meanings. For example, *kalasid* and *kalu* are both plural partitive forms of *kala* 'fish', with the same meaning, and *jalgadel*, *jalul* and *jalgel*, are all adessive plurals of *jalg* 'foot', but vary slightly in meaning; *jalgadel* has an external locational meaning (something is *on the feet*), whereas *jalul* and *jalgel* translate as 'back on the feet'. Nevertheless, regular infected forms make up the majority.

Since irregularity does not provide a full explanation of the present frequency and paradigm effects, the question remains of how to account for these effects in current models of the mental lexicon. One possibility is offered by dual-route models, which hold that forms are stored but rules are also available (Baayen et al., 1997). If so, human memory is capable of storing much more information than previously assumed (Hankamer, 1989; Niemi et al., 1994; Yang, 2016). Alternatively, we may ask whether it is fruitful to discuss these issues in terms of rules and representations. In learning-driven computational models of lexical processing (Baayen et al., 2011; Seidenberg & Gonnerman, 2000) frequency effects, as well as paradigmatic effects, can arise without representations for whole words. Thus, frequency effects for Estonian inflected forms, just as frequency effects for English word n-grams, require that we rethink the traditional Bloomfieldian division of labour between storage and computation.

Chapter 3

Production of Estonian case-inflected nouns shows whole-word frequency and paradigmatic effects

Abstract

Most psycholinguistic models of lexical processing assume that the comprehension and production of inflected forms is mediated by morphemic constituents. Several more recent studies, however, have challenged this assumption by providing empirical evidence that information about individual inflected forms and their paradigmatic relations is available in long-term memory (Baayen et al., 1997; Milin et al., 2009; Moscoso del Prado Martín et al., 2004). Here, we investigate how whole-word frequency, inflectional paradigm size and morphological family size affect production latencies and articulation durations when subjects are asked to read aloud isolated Estonian case-inflected nouns. In Experiment 1, we observed that words with larger morphological family elicited shorter speech onset latencies, and that forms with higher whole-word frequency had shorter acoustic durations. Experiment 2, for which we increased statistical power by using 2,800 words, revealed that higher whole-word frequency, inflectional paradigm size, and morphological family size reduced both speech onset times and acoustic durations. These results extend our knowledge of morphological processing in three ways. First, whole-word frequency effects of inflected forms in morphologically rich languages are not restricted to a small number of very high-frequency forms, contrary to previous claims (Hankamer, 1989; Niemi et al., 1994; Yang, 2016). Second, we

replicated the morphological family size effect in a new domain, the acoustic durations of inflected forms. Third, we showed that a novel paradigmatic measure, inflectional paradigm size, predicts word naming latencies and acoustic durations. These results fit well with Word and Paradigm morphology (Blevins, 2016) and argue against strictly (de)compositional models of lexical processing.

3.1 Introduction

The post-Bloomfieldian tradition of American structuralism builds on the morpheme as the smallest meaningful component of complex words. Although the morpheme still plays a central role in Distributed Morphology (Halle & Marantz 1993, see also Yang 2016), most other theories of morphology have moved away from this theoretical construct. For example, Paradigm Function morphology (Stump, 2001) focuses on how bundles of morphosyntactic features are realized in a certain cell of a paradigmatic class. Word and Paradigm morphology (Blevins, 2003, 2013, 2016; Matthews, 1974) fills paradigm cells by using propositional analogy between inflected forms functioning as principal parts. In these theories, morphemes have no independent theoretical status.

By contrast, most psycholinguistic theories still follow the post-Bloomfieldian tradition and assume that morphemes are psychologically real and essential to understanding and producing complex words (Levelt, Roelofs & Meyer, 1999; Marcus et al., 1995; Pinker, 1999; Taft & Forster, 1976b; Yang, 2016). Some studies hold open the possibility that high frequency forms might be stored along with irregular forms (Marantz, 2013; Rastle et al., 2004). To the extent that units for regular complex words are allowed into the theory, these units only play a role late in the comprehension process. Initial processing is driven by morphemes, late processing allows for whole-word units to contribute as well (Fruchter & Marantz, 2015; Taft, 2004). In particular, it has been claimed repeatedly (Hankamer, 1989; Niemi et al., 1994; Yang, 2016) that storage of inflected words would be computationally inefficient for languages with rich inflectional morphology.

In computational linguistics, morphemes have played a crucial role in older systems for both Finnish (Karlsson & Koskenniemi, 1985; Koskenniemi, 1984) and Estonian (Kaalep, 1997), whereas in newer approaches, its role is much reduced (see e.g., Cotterell, Kirov, Sylak-Glassman, Walther, Vylomova, Xia, Faruqui, Kübler, Yarowsky, Eisner & Mans 2017 and the references therein). Moreover, experimental evidence challenging the post-Bloomfieldian perspective on morphological processing is accumulating. This research shows that whole-word frequency (Baayen et al., 1997; Balling & Baayen, 2008; Caselli et al., 2016; Kuperman et al., 2009), inflectional entropy (Milin et al., 2009; Moscoso del Prado Martín et al., 2004; Tabak et al., 2010) and morphological family size (De Jong et al., 2002; Moscoso del Prado Martín et al., 2004; Schreuder & Baayen, 1997) co-determine lexical processing costs.

Importantly for the present study, Lõo, Järvikivi & Baayen (2017) showed that lexical decision latencies to inflected words in the morphologically complex Finno-Ugric language Estonian were best predicted by both whole-word frequency and inflectional paradigm size, a novel measure counting the number of inflected forms in a given paradigm that people encounter, calculated on the basis of a large corpus. The present study was designed to replicate this finding in production, using a word naming task. We will report in two experiments that production latencies and acoustic durations of Estonian words are facilitated by whole-word frequency, inflectional paradigm size and morphological family size. In what follows, we will first discuss the growing literature on whole-word frequency and paradigmatic effects in other, mostly morphologically less rich languages such as English and Dutch.

3.1.1 Whole-word frequency effects

More frequent regular complex words are recognised and produced faster than less frequent regular words. This whole-word frequency effect for regular complex words emerges independently from the frequencies of individual morphemes in a complex word. Baayen et al. (1997) conducted a lexical decision task with Dutch singular and plural nouns. They found that plural dominant forms (e.g., *eyes*) were recognised faster than singular dominant forms (e.g., *noses*), even when the frequency of the lemma was held constant. Baayen et al. (2003) replicated this finding using auditory lexical decision. Whole-word frequency effects in auditory comprehension have been also reported for Danish (Balling & Baayen, 2008, 2012). Pham & Baayen (2015) observed whole-word frequency effects for Vietnamese compounds, but anti-frequency effects of compounds' constituents.

Some studies have suggested that whole-word frequency effects are in some way restricted. A lexical decision study by Alegre & Gordon (1999) reported that more frequent English inflected nouns and verbs were recognized faster than their infrequent counterparts, however only in the high-frequency range. Similarly, Laine, Niemi, Koivuselkä-Sallinen & Hyönä (1995); Laine et al. (1999); Lehtonen & Laine (2003); Niemi et al. (1994); Soveri et al. (2007) argued for Finnish that whole-word processing occurs only in the highest frequency range of inflected forms. However, Baayen et al. (2007) showed on the basis of a large-scale analysis on the visual lexical decision latencies of over 8,000 morphologically complex words in the English Lexicon Project (Balota et al., 2004) that the whole-word frequency of complex words was an important predictor of processing time across the complete frequency span.

The whole-word frequency effect has been investigated not only in word recognition, but also in production studies. Roelofs (1996) studied the production of complex words with implicit priming task, where participants first learned to associate word pairs and then to produce the second word pair, which was a Dutch nominal compound. He found that speech onset times are sensitive to constituent frequency but not to whole compound frequency. Bien et al. (2005) completed a similar associative production task with Dutch compounds. In this task, participants learned to first associate a compound with a certain visual marker, and were then asked to produce the compound when only this marker was presented on a computer screen. They also found strong constituent effects, but also observed a small nonlinear whole-word frequency.

This result was not replicated in a different paradigm by Janssen, Bi & Caramazza (2008), who studied Chinese and English compounds with a picture naming task. For both languages, more frequent compounds triggered faster responses, but not the constituents. Tabak et al. (2010) studied Dutch inflected verbs with the same task. A facilitatory effect of frequency was observed when photographs with different verbs tenses were presented to the participants. Their study, however, did not provide consistent support for whole-word frequency effects in Dutch. Likewise, another picture naming study with Dutch singular and plural nouns (Baayen, Levelt, Schreuder & Ernestus, 2008) did not find whole-word frequency effects, which is consistent with Levelt et al. (1999) (but see below for paradigmatic effects in this study). In summary, whereas some studies have found the effect of whole-word frequency in production latencies, others have failed to find such an effect (see Janssen et al. 2008 for possible reasons for this discrepancy).

The evidence from studies measuring the acoustic duration of complex words seems to be more conclusive. Pham (2014) conducted a word naming task with 14,000 Vietnamese two-syllable compound words. In this one participant megastudy, more frequent compounds were read aloud faster. However, as this task also requires reading the stimulus, it is not clear to what extent speech onset latencies are revealing about the production process itself. Pham (2014) therefore also investigated the acoustic durations of the words produced, and reported that more frequent compounds were uttered with shorter acoustic durations. Effects of constituent frequency on duration could not be established (see also Sun 2016 for Chinese). Furthermore, Caselli et al. (2016) found in a large-scale study of English conversational speech that the whole-word frequency of English inflected forms correlates negatively with the acoustic duration of the whole inflected form. Finally, Tomaschek & Baayen (2017) observed using electromagnetic articulography more co-articulation between the stem and inflectional affix in more frequent German verbs (see also Tomaschek, Tucker, Wieling & Baayen 2017; Tomaschek et al. 2013). In summary, evidence is accumulating that also for speech production, whole-word frequency plays an important role.

This body of experimental findings suggests that the post-Bloomfieldian model of morphology does not carry over to lexical processing. Nonetheless, a modified version of the post-Bloomfieldian perspective has been put forward for comprehension. According to Fruchter & Marantz (2015); Taft (2004), the initial stages of comprehension are driven by morpheme-based processes, whereas whole-word frequency effects arise at later processing stages when morphemes are combined. However, these predictions about the time-course of comprehension have also been challenged. For instance, Kuperman et al. (2009) conducted an eye-tracking experiment with isolated Dutch compounds, and observed that the frequency of the whole compound facilitated the processing as early on as during the first fixation, when the complete (8-12 characters long) compound had not yet been read. Hendrix (2015) likewise observed whole-word frequency effects in the first fixations when English compounds were read in natural text. An early temporal locus of the whole-word frequency effect was further supported by Schmidtke et al. (2017) in a distributional survival analysis of the lexical decision times, taken from the British Lexicon Project (Keuleers, Lacey, Rastle & Brysbaert, 2012) and the Dutch Lexicon Project (Keuleers, Diependaele & Brysbaert, 2010). They observed that the whole-word frequency effect emerges much earlier in time than constituent frequency effects. Whereas initial obligatory decomposition is challenged by their finding, it does not rule out that word processing might be a function of both constituent and whole-word frequency effects. Note that there are three possible scenarios of early versus late frequency effects, (1) whole-word frequency effects are present in short reaction times, morpheme frequency effects in long reaction times, (2) whole-word frequency effects are in long reaction times, morpheme frequency effects in short reaction times, and (3) whole-word and morpheme frequency effects emerge simultaneously. Whereas the first scenario can only be accommodated in this framework by means of ancillary assumptions explaining why morphemic effects are visible only when processing is least efficient, the second one is straightforwardly compatible with theories incorporating obligatory early decomposition.

3.1.2 Paradigmatic effects

A further problem for strictly decompositional models is that they cannot straightforwardly account for effects that seem to involve not only properties of individual words but also properties pertaining to the paradigms in which they occur. Inflectional entropy measures take into account the probability distribution of forms in a paradigm. High entropy reflects the fact that individual inflected forms within the same inflectional paradigm have a more or less equal frequency; low entropy results if the frequency distribution is far from uniform within the paradigm. In a series of lexical decision experiments with Serbian, English and Dutch complex words, inflectional entropy correlated negatively with response latencies (Baayen et al., 2007; Moscoso del Prado Martín et al., 2004; Tabak, Schreuder & Baayen, 2005). In production, Baayen et al. (2008) Tabak et al. (2010) found an opposite effect in a Dutch picture naming: higher entropy correlated positively with response times (see also Bien, Baayen & Levelt 2011).

Baayen et al. (2011); Milin et al. (2009) investigated a further entropy measure, relative entropy (also known as Kullback-Leibler divergence) with a Serbian lexical decision task. Relative entropy quantifies the difference between the probability distribution of a word's specific paradigm and the probability distribution of its inflectional class. They found that inflected forms from a more typical paradigm were recognised faster.

Further, paradigmatic effects have been reported not only for inflected forms, but also for derived and compound words. Morphological family size, the number of derived and compound words sharing a constituent, e.g., *worker*, *handwork*, *workforce*, shows a negative correlation with response times. Words with more family members are recognised faster. This effect has been documented for many typologically different languages, such as Dutch (Schreuder & Baayen, 1997), English (De Jong et al., 2002), Hebrew (Moscoso del Prado Martín et al., 2005) and Finnish (Moscoso del Prado Martín et al., 2004). The morphological family size effect is taken to be semantic in nature. Words with more family members are a part of a larger network of words with similar meaning, and hence appear to receive more activation from their family (De Jong, 2002). This interpretation is further confirmed by the finding that semantically unrelated morphological family members actually inhibit processing (Moscoso del Prado Martín et al., 2004; Mulder et al., 2014).

Furthermore, a few studies have looked at how paradigmatic relations influence acoustic realizations of complex words. For example, Hay (2001) found that the higher ratio between whole-word and stem frequency in English derived words led to a higher likelihood for the boundary between the stem and affix to get reduced. This was however not replicated by Hanique & Ernestus (2012). Kuperman, Pluymaekers, Ernestus & Baayen (2006) investigated interfixes in Dutch compounds, and reported that the duration of interfixes is dependent on the frequency distribution of the paradigm it belongs to, more probable interfixes have longer durations. This leads to the conclusion that not only individual forms but also their paradigmatic organization matter in production.

Finally, an Italian lexical decision study by Traficante & Burani (2003) looked at how inflectional paradigm size, the number of paradigm members related to a given inflected word, is reflected in the processing costs. For instance, for English *work*, the inflectional paradigm contains the words *work*, *works*, *worked* and *working*. They compared verbs and adjectives and found an inhibitory effect of inflectional paradigm size. Adjectives, which have fewer paradigm members, were recognized faster than verbs, which have more paradigm members. However, their effect of inflectional paradigm size might be an effect of the word category instead. In addition, their study presumed that all members of an inflectional paradigm have the same status in language processing. Yet, in a highly inflected language this is usually not the case, not all forms are usually present.

In contrast, Lõo et al. (2017) used a new measure, the number of actually occurring paradigm members for a given word. Inflectional paradigm size counts were based on the number of inflected forms in use, available in the 15-million token Balanced Corpus of Estonian [1], rather than on the number of forms an Estonian noun paradigm has in principle. Thus, the actual number of paradigm members that people encounter in language use can vary for individual words.

First, not all forms of an inflectional paradigm are necessarily realized in actual use. Most Estonian nouns are not used in in all their 14 singular and plural forms, but only in cases which make sense based on the meaning of the word. This point

[1] http://www.cl.ut.ee/korpused/grammatikakorpus/(15.04.2017)

is illustrated for Finnish by Karlsson (1986)'s corpus-based survey. He shows that the number of forms available for speakers of Finnish depends on the semantics of the word. For example, the word *kesä* 'summer' has mostly a temporal meaning, thus the word frequently occurs in adessive case (e.g., *kesällä* 'in the summer'), but less often or not at all in many other cases. Likewise in Estonian, the number of inflected forms in use varies from word to word. For example, essive case (e.g., *jalana* 'as foot/leg') is usually not used with the word *jalg* 'foot; leg'. However, at the same time, some paradigms may have multiple members for a given slot. For example, *jalgadel*, *jalul*, *jalgel* are all plural adessive forms of *jalg* in use.

Using this corpus-based count and looking at the paradigm size for Estonian nouns, a richly inflecting language, Lõo et al. (2017) found a facilitatory effect of paradigm size in both a lexical decision and a semantic categorization task. The fact that an effect of inflectional paradigm size emerged also in a task where participants were asked to determine whether a word on the screen refers to an animate or inanimate entity suggests that similar to the effect of the morphological family size, the inflectional paradigm size effect might be semantic in nature.

In summary, previous studies indicate that highly specific paradigmatic properties of complex words co-determine how we recognise, read and produce such words. The whole-word frequency effect and various paradigmatic effects emerge across modalities, tasks and languages. Most of the evidence, however, coming from languages such as Dutch and English, which have relatively simple morphology, provides limited possibilities to study the role of paradigmatic relations.

3.1.3 The current study

In the current study, we investigated the effect of whole-word frequency, inflectional paradigm size and morphological family size in a morphologically rich Finno-Ugric language, Estonian, whose nominal paradigms are characterized by 14 cases in both singular and plural. Considering that many cases also have overabundant forms with subtle meaning differences, the total number of forms for a single noun lemma can theoretically be well over 40 (see Table 3.1 for *jalg* 'foot').

In addition to having a complex inflectional system, Estonian is characterized

Table 3.1: Inflectional paradigm of jalg 'foot, leg' with 46 members.

Case	Singular	Plural
Nominative	jalg	jalad
Genitive	jala	jalgade, jalge
Partitive	jalga	jalgasid, jalgu
Illative-1	jalga	-
Illative-2	jalasse	jalgadesse, jalusse, jalgesse
Inessive	jalas	jalgades, jalus, jalges
Elative	jalast	jalgadest, jalust, jalgest
Allative	jalale	jalgadele, jalule, jalgele
Adessive	jalal	jalgadel, jalul, jalgel
Ablative	jalalt	jalgadelt, jalult, jalgelt
Translative	jalaks	jalgadeks, jaluks, jalgeks
Terminative	jalani	jalgadeni, jalgeni
Essive	jalana	jalgadena
Abessive	jalata	jalgadeta
Comitative	jalaga	jalgadega

by productive derivation and compounding. Like Finnish (Moscoso del Prado Martín et al., 2004), most Estonian words are part of large morphological families with up to a 1,000 members. Table 3.2 presents a fragment of the Estonian morphological family for *jalg* 'foot, leg', which has over 500 members. There are family members where English translations contain the stem *jalg* 'foot', e.g., *lampjalg* 'flatfoot'. However, many English translation equivalents do not contain the words foot or leg, e.g., *sõnajalg* 'fern'. Moscoso del Prado Martín et al. (2004) made a similar observation for Finnish, where kirja ('book') is found in derived words and compounds such as *kirjepaino* ('paper weight'), *kirjailijantoiminta* ('authorship'), and *kirjoituskone* ('typewriter'). They hypothesized that complex words that are not immediate descendants of a given stem are at semantic distances that are too great for a morphological family size to be present, and reported experimental evidence that this is indeed the case. In the present study, only immediate morphological descendants were used as stimuli.

Whereas Lõo et al. (2017) found whole-word effects in word comprehension, the current study investigates whether they also persist in word production. We consider both the production latencies and the acoustic durations of the words produced. The production latencies reflect both reading processes and processes

Table 3.2: Part of morphological family for jalg 'foot, leg'.

Family member	Meaning	Family member	Meaning
jalgsi	afoot	lampjalg	flatfoot
jalam	base	käskjalg	courier
jalats	footwear	lülijalgne	arthropod
jalamatt	doormat	raskejalgne	pregnant
jalutama	to walk	sõnajalg	fern
jalamaid	instantly	jooksujalu	fast
jalgratas	bicycle	küünlajalg	candleholder
jalgpall	football	rahujalal	in peace
jalgtee	footpath	varesejalad	bad handwriting
jalgpidur	footbrake	puujalg	peg

leading up to the initiation of articulation. The acoustic durations provide a record of how the production of the word unfolded over time. We present two word naming experiments. Experiment 1 is a small-scale study in which each participant read aloud the same 200 case-inflected nouns. Experiment 2 implemented a large-scale design in which each participant read aloud 400 case-inflected nouns from an item list with in total 2,800 nouns. In the discussion section, we discuss the results and how they relate to theories of morphological processing.

3.2 Experiment 1: Small-scale word naming study

3.2.1 Participants

26 native speakers of Estonian (18 female; age 21-67, mean 38.66, sd 14.91 years) with normal or corrected-to-normal vision were recruited from Tallinn University in Estonia. They received 10 euros for their participation.

3.2.2 Materials

200 case-inflected nouns, 100 animate and 100 inanimate, were selected from the Balanced Corpus of Estonian. Whole-frequency of the items ranked between 1 and 213 per million (median 5). Lemma frequency ranked between 1 and 3402 per million (median 892). Compounds were not included in the data set. Inflected forms had either a simplex (e.g., *maja+s* 'in the house') or a complex stem, which consists

of a root and a derivational ending (e.g., *raha+stuse+d* 'fundings'). As more than half of the inflected forms in the Estonian corpus have complex stems, materials were selected such that complex stems would be represented (roughly 50%). The length in letters varied between 4 and 15 characters (median 8 characters).

3.2.3 Apparatus

The experiment was conducted in a sound attenuated room. Responses were recorded using a Korg MR-1000 recorder and a Countryman ISOMAX earset microphone. The experiment was programmed in ExperimentBuilder by SR Research Ltd. Speech onset latency and acoustic duration measures were retrieved using a Matlab script. The stimuli were presented on a 21-inch Dell computer screen in lower case 26-point Courier New Bold font, using the ExperimentBuilder software by SR Research Ltd.

3.2.4 Procedure

Participants were asked to read aloud words on the computer screen as naturally as possible. Trials started with a blank screen for 1,000 ms, replaced by the fixation cross for 500 ms, after which the stimulus appeared on the screen for 1,500 ms in the middle of the screen. The experiment started with six practise trials which were followed by 200 experimental trials. The same 200 items were presented to all participants in a randomized order. The task took approximately 25 minutes to complete. In addition, each participant also completed a semantic categorization experiment with the same set of items. The results of the semantic categorization experiment are discussed in Lõo et al. (2017). Whether participants started with the naming or categorization task was counterbalanced.

3.2.5 Predictors

We are interested in the following frequency and paradigmatic predictors: (1) *Whole-word frequency* captures the total number of occurrences of a particular form in a corpus. In case of syncretic forms (e.g., the form *viilu* represents genitive, and partitive singular of 'slice'), we took the frequency of the most frequent inflectional case. As Estonian displays a three way phonemic length distinction, and vowel or consonant in stressed syllables can be pronounced either short, long or extra long, ambiguity is often resolved in pronunciation. For example, the written form *viilu* can either be pronounced with a long *i* as in [vi:lu] (genitive) or with an extra long *i* as in [vi::lu] (partitive). In particular, in a production task, the participant is likely to make a choice in favour of a specific case. The most frequent case function is usually the dominant one. When a different decision is made, and frequencies are accumulated across orthographically identical forms, the whole-word frequency effect becomes even stronger (without changing the pattern of results). (2) *Lemma frequency* is the cumulative frequency of a complete inflectional paradigm.

(3) *Inflectional paradigm size* is the number of observed forms for a certain lemma. We excluded orthographic neighbours from the inflectional paradigm size to ensure that the effect was not confounded with orthographic neighbourhood density. For example, when we calculated inflectional paradigm size for *jalata*, we excluded forms such as *jalana* and *jalaga* as they are also orthographic neighbours. However, using a paradigm size count including orthographic neighbours led to results that are very similar to those reported below. Furthermore, we have chosen to go with the linguistic characterization of paradigms, giving giving full recognition to overabundant and syncretic forms. As mentioned above, syncretic forms are often pronounced differently and have a separate morphosyntactic function (see also Plag, Homann & Kunter 2015 for systematic differences in the duration of the [s] in English, depending on its morphological function). Alternative analyses are again possible, for instance, an analysis in which syncretic forms are collapsed.

(4) *Inflectional entropy* measures the average amount of information in an inflectional paradigm. The higher the entropy, the more uniformly the inflected forms are distributed and as a result the higher the uncertainty. The inflectional entropy (*HI*) was calculated using a formula adapted from Moscoso del Prado Martín et al. (2004):

$$\mathrm{HI}(w) = -\sum_{i=1}^{n} p(w_i)\log_2 p(w_i) = -\sum_{i=1}^{n} \frac{f(w_i)}{f(w)}\log_2\frac{f(w_i)}{f(w)}$$

where *w* is the lemma, *n* the number of inflected variants, *f(w)* lemma frequency, the summed frequency of variants,*f(w_i)* is the frequency of an particular invariant; and *p(w_i)* is the probability within the paradigm. All four measures were based on the 15-million token Balanced Corpus of Estonian. Forms with zero frequency were not included when calculating HI(w). Finally, (5) *morphological family size* is the number of compound and derived words which share the target word as a constituent (e.g., *jalg* is part of *jalgsi*, *jalam*, *jalats*, see Table 3.2) This measure was calculated using an online version of the Estonian Word Families dictionary (Vare, 2012).

In addition to these predictors, we added several variables as controls. (6) *Orthographic length* (the number of characters) is a rough approximation of length in phonemes, as Estonian has a shallow orthography. The number of graphemes closely corresponds to the number of phonemes. (7) *Orthographic neighbourhood density* is the count of words that differ only by one letter. This measure was obtained with the function *levenshtein.distance* from the R-package *vwr* (Keuleers, 2013). We excluded orthographic neighbours from the same inflectional paradigm from this count. Further, we included (8) *nominal case* (three syntactic cases and eleven semantic cases), (9) *experiment order* (whether naming or semantic categorization task was first) and (10) *stem complexity* (simple or complex) as control variables. Finally, two phonetic variables were added: (11) *manner of articulation* (five levels: *approximant*, *fricative*, *nasal*, *plosive*, *trill*), and (12) *place of articulation* (five levels: *labial*, *alveolar*, *palatal*, *velar*, *glotta*l) of the first segment of the word.

3.2.6 Analysis and Results

Prior to the analysis, the data was cleaned and transformed (Baayen & Milin, 2010). First, production latency outliers, i.e., responses longer than 1,500 ms (0.5% of the data) were removed. Second, we excluded acoustic durations shorter than 200 ms and longer than 2,500 ms (1.5% of the data). Finally, eight stimuli which belonged to more than one inflectional paradigm (4% of the data) were also excluded.

In order to reduce the skewness in the distribution due to outliers, production latencies, acoustic durations, whole-word frequency, lemma frequency and morphological family size were log-transformed. In light of the high collinearity between predictors (see Table 3.3 in Appendix A), the effects of our key predictors were tested both together and separately in order to clarify whether results in the joint model were affected by enhancement or suppression.

We analysed the data with the Generalized Additive Mixed Models (GAMM, Wood 2006; R-package *mgcv*, see also Baayen et al. 2017). A standard Gaussian regression model predicts responses y_i as a function of a linear predictor η_i and and error term. The linear predictor is a weighted sum of an intercept β_0 and one or more predictors, for instance.

$$y_i = \eta_i + \epsilon_i \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2) \text{ and } \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

The generalized additive model (Hastie & Tibshirani, 1990; Lin & Zhang, 1999; Wood, 2006, 2011; Wood, Goude & Shaw, 2015) extends the linear predictor with one or more terms that are functions of one or more predictor variables. In the model

$$y_i = \beta_0 + \beta_1 x_{1i} + f(x_{2i}) + \epsilon_i, \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2).$$

the effect of the second predictor x_2 on the response is modulated by a function $f()$ that is optimized to detect and evaluate non-linear functional relations between x_2 and y. If the functional relation between x_2 and y is indeed nonlinear, the nonlinear trend will be modeled as a smooth that is weighted sum of (simple nonlinear) basis functions. If there is no nonlinearity, $f()$ reduces to a straight line. In order to balance faithfulness to the data against model parsimony, smooths are penalized for wiggliness. The effective degrees of freedom (edf) of a smooth, which are used to evaluate the significance of a smooth, reflect the degree of penalization. Penalization may result in all wiggliness being removed from the smooth, resulting in a term with one effective degree of freedom, in which case the effect of the predictor is linear. Nonlinear terms in the model are interpreted by plotting the partial effect of the smooth together with a 95% confidence interval. The generalized additive mixed model incorporates random-effect factors, which are modeled by functions that impose a ridge penalty. This ridge penalty makes it expensive for random-effect coefficients to have large values. As a consequence, coefficients are shrunk towards zero, as in the linear mixed model. The summary of a GAMM reports both the parametric part of the model (intercept and the betas of the linear terms) and the smooths (wiggly curves and wiggly (hyper)surfaces, as well as random effects).

The residuals of the models we initially fitted to the data showed a departure from normality with heavy tails, characteristic of the t-distribution, we made use of the the scaled *t*-distributed family. The statistical models presented below are the result of exploratory data analysis. A backward stepwise modelling procedure was followed in which insignificant predictors were removed one by one. Predictors were initially modeled as nonlinear effects, but whenever support for nonlinearity was not granted, they were entered as linear terms into the model specification. For additional model comparisons and visualizations, we made use of the package *itsadug*. By-subject factor smooths for trial were sufficient to remove autocorrelations from the residual error.

Production latencies

A GAMM fitted to log-transformed production latencies revealed that response times increased linearly with orthographic length and orthographic neighbourhood density, and decreased linearly with morphological family size. Inflectional paradigm and whole-word frequency were not significant in this model, possibly due to a high correlation between the predictors in the experiment (see Table 3.3). For both experiments, neither order nor stem complexity was predictive of production latencies. They were tested both as main effects and in interactions with other predictors. The model included by-participant factor smooths for trial, as well as by-participant random slopes for length, it also included random intercepts for item and manner of articulation. Further main or random effects did not reach significance. The complete model summary can be found in Table 3.3.

Acoustic durations

Acoustic durations decreased linearly for increasing whole-word frequency (see Figure 3.1), and orthographic neighbourhood density. Neither experiment order nor stem complexity was again predictive of acoustic durations. They were tested

Table 3.3: Summary of the partial effects in GAMM fitted to log-transformed production latency in Experiment 1.

A. parametric coefficients	Estimate	Std. Error	z-value	p-value
(Intercept)	0.39	0.01	27.24	< 0.0001
Morphological family size	-0.004	0.001	-5.17	< 0.0001
Orthographic neighbourhood density	0.01	0.002	3.67	0.0002
Orthographic length	0.01	0.001	9.12	< 0.0001
B. smooth terms	edf	Ref.df	Chi.sq-value	p-value
s(Orthographic length,Participant)	23.51	25.00	401.14	< 0.0001
s(Trial,Participant)	155.18	233.00	11081.69	< 0.0001
s(Item)	126.73	183.00	452.04	< 0.0001
s(Manner)	4.51	5.00	972.59	< 0.0001

both as main effects and in interactions with other predictors. Orthographic length was not included in the model as length in letters in not a predictor (or cause) of acoustic duration and its strong correlation reflects ($r = 0.75$) that it may simply be the written counterpart of acoustic duration. Inflectional paradigm size and morphological family size were not significant predictors in the model. The model included by-participant factor smooths for trial, as well as random intercepts for item and nominal case. Further random effects did not reach significance. The complete model summary can be found in Table 3.4.

Table 3.4: Summary of the partial effects in GAMM fitted to log-transformed acoustic duration in Experiment 1.

A. parametric coefficients	Estimate	Std. Error	z-value	p-value
(Intercept)	0.56	0.01	40.94	< 0.0001
Whole-word frequency	-0.01	0.003	-4.26	< 0.0001
Orthographic neighbourhood density	-0.06	0.01	-9.28	< 0.0001
B. smooth terms	edf	Ref.df	Chi.sq.-value	p-value
s(Trial,Participant)	149.18	233.00	15434.41	< 0.0001
s(Item)	174.34	184.00	27613.32	< 0.0001
s(Case)	7.70	14.00	61594.68	0.0001

In summary, production latencies decreased with increasing morphological family size, and acoustic durations decreased with increasing whole-word frequency. The high correlations between critical predictors in the materials of this experiment, as well as the relatively small number of items, led us to design Experiment 2. Experiment 2 increases statistical power by expanding the number of different

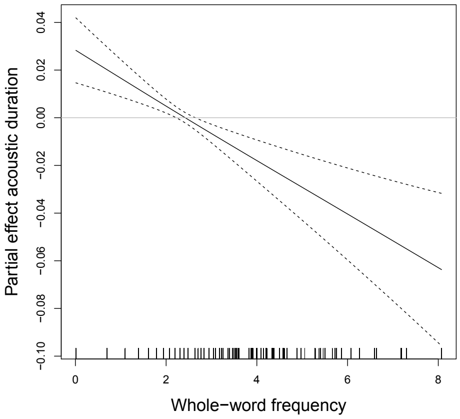

Figure 3.1: Partial effects for whole-word frequency for acoustic duration of Experiment 1.The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for whole-word frequency.

words to 2,800. Furthermore, by including more items, we were able to substantially reduce the correlation between inflectional paradigm size and morphological family size.

3.3 Experiment 2: Large-scale word naming study

3.3.1 Participants

Thirty-three native speakers of Estonian (20 females; age 22-69, mean 38.34, sd 15.09 years) with normal or corrected-to normal vision and no speech impairments participated in the study. They were tested at University of Tartu and at Tallinn University of Technology in Estonia. Data for one participant was removed from the analysis, as he did not complete the experiment due to technical difficulties. Participants received 15 euros for their participation.

3.3.2 Materials

A total of 2,800 case-inflected nouns were selected from the Balanced Corpus of Estonian. The frequency distribution of the items ranged between 1 and 1000 per million (median 40). Lemma frequency ranged between 1 and 3439 per million (median 519). Compounds were excluded from the data set. However, about 40%

of the items had a complex stem consisting of a root and a derivational ending. Also in this experiment, stimuli had both simplex and complex stems to resembled the distribution in our corpus.The length of stimuli varied between 2 and 19 characters (median 8 characters). Stimuli were divided over 28 experimental lists, each containing 400 items. An overlapping design was used with a 300-word overlap between successive lists. Each stimulus elicited four responses.

3.3.3 Apparatus

The experiment was conducted in a sound attenuated booth. Participants' responses were recorded with a Marantz PMD670 digital recorder, using a supercarioid condenser table top microphone by Beyerdynamic, placed approximately 10 cm from participants mouth. Participants were also wearing an EyeLink II head-mounted eye tracker by SR Research, which recorded their eye movements. The eye-tracking data is still under analysis and not reported here. The stimuli were presented on a 21-inch Dell computer screen in lower case 26-point Courier New Bold font, using the ExperimentBuilder software by the same company.

3.3.4 Procedure

Participants were instructed to read aloud single words appearing on the computer screen as naturally as possible. Each trial started with a drift correction on the left of the screen, after which the target appeared in the center. The target stayed on the screen for 1,500 ms and was then replaced by a fixation cross that remained on the screen for 2,500 ms. The experiment started with ten practice trials, which were followed by 400 experimental trials. Every 100th trial was followed by a short break. The break lasted until the participant indicated he was ready to continue. At the end of the experiment, participants filled out a language background questionnaire. Participants were asked to rate their foreign and native language proficiency and use, e.g., the number of languages they speak, the number of books they had read in the past month and how communicative they are. The whole procedure lasted approximately 90 minutes.

3.3.5 Analysis and Results

Prior to the analysis, we removed outliers from the data set. First, production latencies shorter than 1,500 ms (3% of the data) were excluded from the dataset. Second, acoustic durations shorter than 200 ms or longer than 2,000 ms (1.6 % of the data) were excluded. Finally, 65 items which belonged to more than one inflectional paradigm (2.3% of the data) were removed from the data set. To reduce the effect of outliers, production latencies, acoustic durations, orthographic neighbourhood density and morphological family size were log-transformed, and inflectional paradigm size was transformed using the power transformation of 0.75 (*powertranform-function* from the R-package *car* by Fox & Weisberg 2011). Table 8 in Appendix A presents correlation coefficients between the predictors of the current experiment. Correlations between most predictors were still substantial, but nevertheless reduced compared to Experiment 1. For instance, the correlation between inflectional paradigm size and morphological family size was only 0.30 in Experiment 2 (0.62 in Experiment 1), which facilitates teasing apart statistically the effects of these two paradigmatic measures.

Production latencies

A GAMM fitted to the log-transformed production latencies showed that increasing morphological family size (see the upper right panel of Figure 3.2) linearly decreased production latencies. Response latencies increased with whole-word frequency. The upper left panel of Figure 3.2 shows that this effect is nonlinear as indicated by the wider confidence intervals the model is less certain at the higher frequency range. Response latencies also decreased nonlinearly with inflectional paradigm size. The upper middle panel of Figure 3.2 shows this effect levels off as the paradigms increase. Finally, orthographic length increased production latencies. The lower left panel of Figure 3.2 shows that this effect is slightly nonlinear for the longer words. Orthographic neighbourhood density did not reach significance as a main effect. Stem complexity was also not predictive of production latencies. It was tested both as a main effect and in interactions with other predictors. The model included by-participant factor smooths for trial, as well as by-participant random slopes for frequency, length, inflectional paradigm size and orthographic neighbourhood density. The model also included random intercepts for item, nominal case, manner of articulation and place of articulation. Further random effects did not reach significance. The complete model summary can be found in Table 3.5.

Table 3.5: Summary of the partial effects in GAMM fitted to log-transformed production latency in Experiment 2.

A. parametric coefficients	Estimate	Std. Error	z-value	p-value
(Intercept)	-0.41	0.03	-12.57	< 0.0001
Morphological family size	-0.004	0.001	-3.52	0.0004
Orthographic neighbourhood density	-0.003	0.003	-0.78	0.43
B. smooth terms	**edf**	**Ref.df**	**Chi.sq.-value**	**p-value**
s(Whole-word frequency)	2.67	3.19	53.13	< 0.0001
s(Inflectional paradigm size)	2.95	3.56	16.01	0.002
s(Orthographic length)	3.19	3.85	72.78	< 0.0001
s(Whole-word frequency, Participant)	8.32	31.00	14.42	0.03
s(Inflectional paradigm size, Participant)	10.55	31.00	18.61	0.02
s(Ort. neighb. density, Participant)	17.63	31.00	87.02	< 0.0001
s(Orthographic length, Participant)	25.07	31.00	363.73	< 0.0001
s(Trial,Participant)	220.06	287.00	812585.72	< 0.0001
s(Item)	573.82	2733.00	747.76	< 0.0001
s(Case)	6.65	14.00	61.82	< 0.0001
s(Manner)	4.55	5.00	1002.03	0.005
s(Place)	3.58	5.00	433.86	0.02

Acoustic durations

Acoustic durations decreased linearly with increasing whole-word frequency (the upper left panel of Figure 3.3), and slightly nonlinearly with inflectional paradigm size (the upper middle panel of Figure 3.3), orthographic neighbourhood density (the lower left panel of Figure 3.3) and nonlinearly with morphological family size (see the upper right panel of Figure 3.3). Stem complexity was not predictive of acoustic durations, neither as a main effect nor in interactions with other predictors. The model also included by-participant random slopes for orthographic neighbourhood density, by-participant factor smooths for trial, and random intercepts for item, nominal-case, and by-manner-of-articulation. Further random effects did not reach significance. The complete model summary can be found in

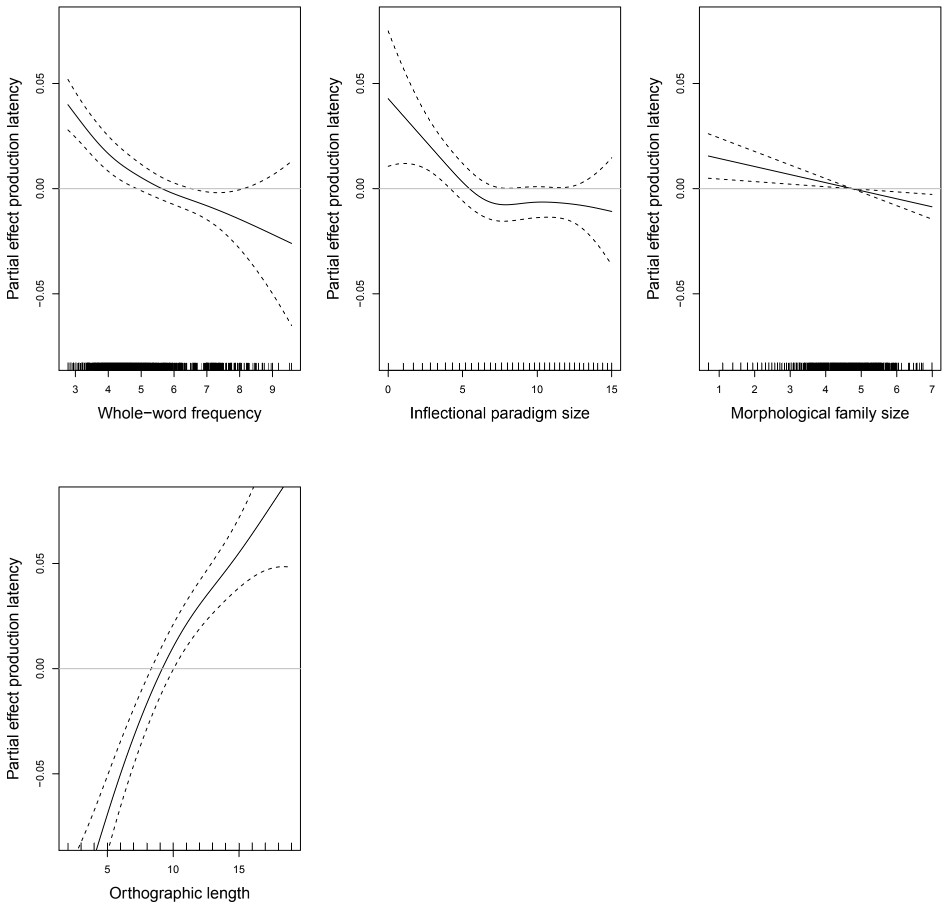

Figure 3.2: Partial effects for whole-word frequency, inflectional paradigm size, morphological family size and orthographic length for the production latencies of Experiment 2. The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

Table 3.6.

Table 3.6: Summary of the partial effects in GAMM fitted to log-transformed acoustic duration in Experiment 2.

A. parametric coefficients	Estimate	Std. Error	z-value	p-value
(Intercept)	0.49	0.02	31.40	< 0.0001
Whole-word frequency	-0.01	0.001	-5.09	< 0.0001
B. smooth terms	edf	Ref.df	Chi.sq.-value	p-value
s(Inflectional paradigm size)	1.54	1.58	16.76	0.0004
s(Morphological family size)	4.01	4.09	13.32	0.01
s(Ort. neighb. density)	2.23	2.28	1003.61	< 0.0001
s(Ort. neighb. density, Participant)	26.39	31.00	662.69	< 0.0001
s(Trial, Participant)	250.01	287.00	93550.45	< 0.0001
s(Item)	2412.41	2669.00	30580.95	< 0.0001
s(Case)	12.95	14.00	73556.81	< 0.0001
s(Manner)	4.33	5.00	7118.09	< 0.0001

Quantile regression analysis of production latencies

The GAMM analysis in Section 3.5.1 indicated that whole-word frequency, inflectional paradigm size and morphological family size co-determine mean production latency in Experiment 2. What a mean (or expected) reaction time cannot tell us, however, is whether a certain variable is already predictive for short responses or whether it is perhaps predictive only for long responses.

Following the general approach of Schmidtke et al. (2017), we investigated the time-course of whole-word frequency, inflectional paradigm size and morphological family size, but instead of using survival analysis, we used quantile regression (R-package *qgam* by Fasiolo et al. 2016). Quantile regression makes it possible to model the relation between a set of predictor variables and a specific percentile of the response variable. We modeled the three predictors of theoretical interest as linear, while including trial and orthographic length as control variables as well as by-participant random intercepts (adding item as a second random effect factor may lead to catastrophic data sparsity at extreme quantiles). Figure 3.4 presents the slopes for the three critical predictors at four deciles (.20, .40, .60 and .80) of the production latency distribution. In all cases slopes were significantly different from zero, even at the earliest deciles. The size of the effects varied across quantiles. For instance, whole-word frequency has the biggest influence already

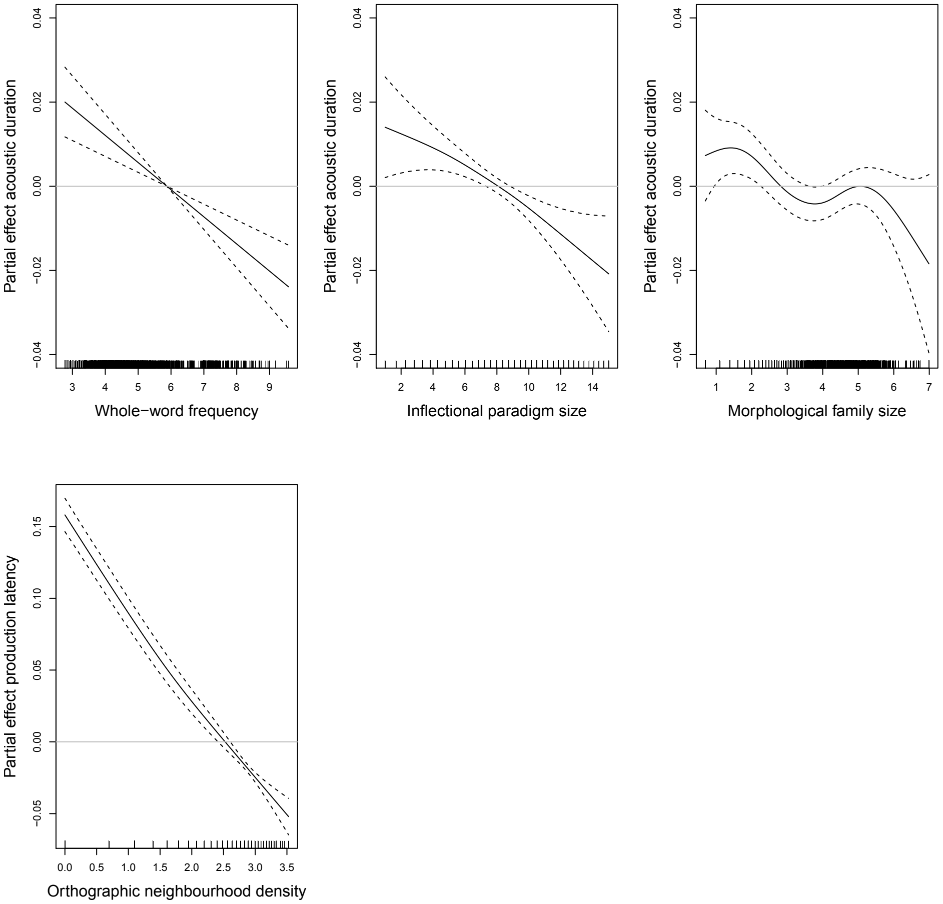

Figure 3.3: Partial effects for whole-word frequency, inflectional paradigm size, morphological family size and orthographic density for acoustic duration of Experiment 2. The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

for short production latencies at the .40 decile (see the left panel of Figure 3.4), whereas inflectional paradigm size and morphological family size have their peaks later on at the .80 decile and .60 decile, respectively (see the middle and right panel of Figure 3.4).

In summary, the quantile regression analysis complements the analysis of mean reaction time in two ways. First, for Estonian inflected forms, the whole-word frequency effect is present already at the early quantiles. This finding is in line with work by Schmidtke et al. (2017), who observed, using survival analysis that the effect of whole-word frequency emerges earlier than the effect of constituent fre-

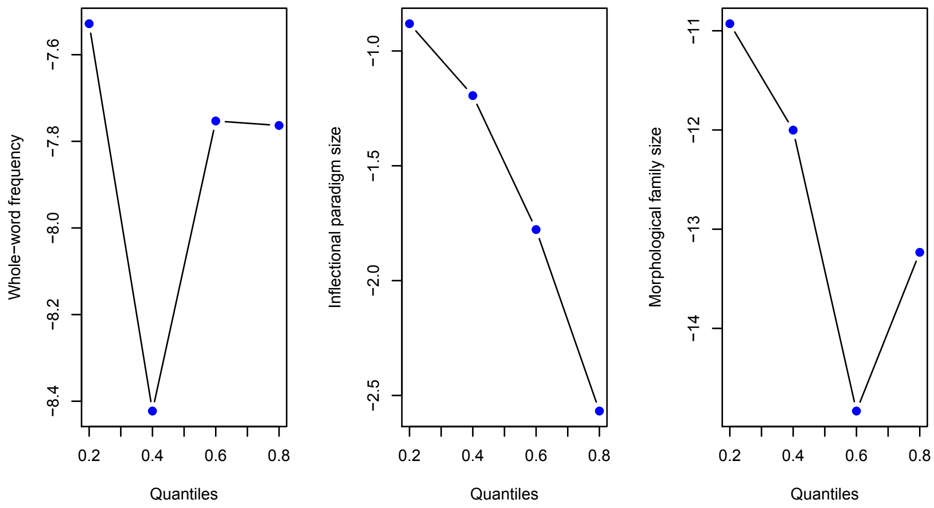

Figure 3.4: Whole-word frequency, inflectional paradigm size, morphological family size β *- coefficients in the second, fourth, sixth and eighth decile of production latency in Experiment 2.*

quency. Second, the quantile regression analysis clarifies that inflectional paradigm effects and morphological family size effects are strongest at later quantiles, which fits with their interpretation as semantic effects (De Jong, 2002; Lõo et al., 2017). Whereas whole-word frequency is a property affecting only a certain inflected form, morphological family size and inflectional paradigm size counts involve the activation of multiple words.

3.4 General Discussion

We have shown that whole-word frequency, morphological family size and inflectional paradigm size predict both processes leading up to the initiation of articulation as well as the processes governing how the production of Estonian case-inflected nouns unfolds over time. In Experiment 1, we showed that words with larger morphological families elicited shorter speech onset latencies, and that forms with higher whole-word frequency had shorter acoustic durations. Experiment 2 revealed that higher whole-word frequency, inflectional paradigm size, and morphological family size reduced both speech onset times and acoustic durations. In the following, we will discuss these results and how they relate to theories of morphological processing.

3.4.1 Paradigmatic effects

It is noteworthy that inflectional entropy was also a significant predictor of production latencies and acoustic durations. However, a simpler measure, inflectional paradigm size outperformed inflectional entropy across analyses. As the correlation between inflectional paradigm size and entropy was quite high in both experiments (Experiment 1: $r = 0.7$; Experiment 2: $r = 0.6$), and paradigm size was a better predictor than entropy, we decided to include only this measure in our analysis. We propose two reasons why this might be the case. First, for small paradigms that are necessarily quite "dense" for high-frequency words (no empty cells), entropy and relative entropy measures are well-defined, whereas for the Estonian paradigms, we are dealing with large and, as outlined above, necessarily more sparse paradigms, to which entropy measures are not easily applied (as one has to make non-trivial decisions as to how to back off from zero probabilities). One option is to increase zero probabilities by some small amount, but what this amount should be is unclear. Another option is to ignore empty cells altogether, but in this case an entropy measure will strongly reflect the number of non-zero cells – recall that for a uniform probability distribution, H = -log(V), with V indicating the number of different (non-zero) probabilities.

The same problem arises for the relative entropy measure. The study of Milin et al. (2009) carefully selected Serbian nouns for which all case forms had frequencies greater than zero. As most Estonian nouns fail this selection criterion, further research is required to determine how to properly back off from zero when calculating relative entropies. Addressing this issue is beyond the scope of the present study. Second, an increasing corpus size will to some extent shift explanatory power from a type count (paradigm size) to one or more measures characterizing the distribution of the token counts of the inflectional variants. Hence, it is conceivable that in substantially larger corpora, measures such as inflectional entropy will outperform the simple paradigm size measure that we use here.

However, for Estonian, inflectional paradigm size is expected to remain a relevant predictor as the number of case forms will not be easily exhausted as for example for German (see Blevins, Milin & Ramscar 2017). Similarly, in English the number of possible preposition+noun combinations is not going to be realized for a given noun in a large enough corpus. For instance, temporal prepositions such as *throughout* do not combine well with static objects such as *telephone* or *chair*. This is noteworthy because many Estonian inflected forms (e.g., *majas*) are functionally similar to English prepositional phrases (e.g., *in the house*). Precisely the richness of the Estonian inflectional system makes it more specific than the Serbian or German system, and this specificity makes it less likely that all paradigm cells will be filled even if the corpus is large enough.

3.4.2 Frequency and memorization

Further, whole-word frequency across the frequency span was a consistent predictor of both production latencies and articulation durations. This finding is at odds with claims that only irregular (Pinker, 1999) or only very frequent (Niemi et al., 1994) inflected forms would show whole-word frequency effects. Furthermore, as suggested by the quantile regression analysis, the effect of whole-word frequency is already detectable at small deciles, indicating it is present in very short reaction times. This result challenges theories positing that the earliest stages of lexical processing are driven by only morphemes (e.g., Fruchter & Marantz, 2015; Taft, 2004). If this were the case, one would expect lemma frequency effects to arise before whole-word frequency effects, contrary to fact.

Interestingly, lemma frequency is also predictive of naming latencies and acoustic durations, but it consistently provided worse model fits. In order to make sure that results are not specific to linear regression, we also analyzed the data using random forests (using the *party* package (Hothorn, Buehlmann, Dudoit, Molinaro & Van Der Laan, 2006) for R). The variable importance of lemma frequency was inferior to that of whole-word frequency for both experiments, across reaction times and acoustic durations. At least for Estonian, lemma frequency does not afford the predictive precision that comes with whole-word frequency.

The support for a whole-word frequency effect for Estonian case-inflected nouns does fit very well with the findings from recent years indicating that in English, sequences of words also show frequency effects. Bannard & Matthews (2008) found in a phrase repetition task that children produced more frequent phrases faster than less frequent phrases (e.g., *a drink of tea* and *a drink of milk*), even when the frequencies of individual words in the sequence were controlled for. Arnon & Snider (2010) found the same frequency effect in adult language processing with a lexical decision task of multi-word phrases. Tremblay et al. (2011) extended the finding to lexical bundles in a series of self-paced reading experiments. More often occurring lexical bundles (i.e., words that occur frequently together, but do not necessarily make up a phrase, e.g., *in the middle of the*) were read faster compared to less frequent controls (e.g., *in the front of the*). Speakers are also able to provide estimates of the frequencies with which word n-grams are used (Shaoul, Baayen & Westbury, 2014; Shaoul, Westbury & Baayen, 2013).

Furthermore, Tremblay & Tucker (2011) had participants produce four-word sequences, and established that participants produced more frequent sequences faster and that acoustic durations were shorter for these sequences. Janssen & Barber (2012) had participants name drawings of adjective and noun pairs, and observed that production latencies decreased with increasing frequency of the multi-word phrase. Sprenger & van Rijn (2013) had participants produce Dutch clock time expressions, and found that more frequent expressions were produced faster. Arnon & Cohen Priva (2013) conducted a multi-word sequence naming experiment and a corpus study of spontaneous speech. They report that acoustic duration was shorter for frequent phrases in both elicited and spontaneous speech, regardless of syntactic boundaries and individual word frequencies.

Interestingly, just as English prepositional phrases such as *in time* and *on foot* can have idiosyncratic, semantically opaque shades of meaning, Estonian case-inflected nouns can have both transparent and idiomatic interpretations. For instance, the form *käes* has a literal meaning 'in the hand', but also an opaque meaning 'due', which cannot be derived compositionally from stem and suffix.

This brings us naturally to the question of whether the whole-word frequency effect is driven solely by such idiomatic case forms? We think this is unlikely. First, a majority of forms have a straightforward transparent interpretation. Second, and more importantly, if semantic irregularity were driving the effect, then it is unclear why an inflectional paradigm size effect is present. The Bloomfieldian lexicon, as the repository of the unpredictable, cannot explain the regular paradigmatic relations that are quantified by the inflectional paradigm size effect.

3.4.3 General remarks

Although Distributed Morphology (Halle & Marantz, 1993) can be taken as a theory of what possible words are and how such words are interpreted, substantial experimental work has been conducted to show that it actually provides an adequate functional characterization of lexical processing (see Fruchter & Marantz, 2015, and references cited there). However, the evidence we report here is not compatible with the processing claims of Distributed Morphology, and further research is required to clarify what gives rise to the very different results obtained by researchers working within Distributed Morphology and researchers working within Word and Paradigm morphology.

Nevertheless, one thing is clear: our results invalidate one line of reasoning that has been widely accepted, and that is exemplified by the work of Hankamer and of Yang. Hankamer (1989)[p.404] argued that

> "A careful examination of morphological complexity in agglutinating languages shows clearly that the full-listing model cannot be an adequate model of general natural-language word recognition. In such languages, parsing must be involved in human word recognition, and not just for rare or unfamiliar forms" (Hankamer, 1989).

More recently, Yang (2010)[p.1168] claimed that " [. . .] the combinatorial explosion of morphologically complex languages necessitates a stage-based architecture of processing that produces morphologically complex forms by rule-like processes [. . .].

At the minimum, the stem must be retrieved from the lexicon and then combined with appropriate rules/morphemes". The form frequency effects and inflectional paradigm size effects for Estonian, and the word n-gram frequency effects for English reviewed above, show that the human brain comprises a memory system that goes far beyond what is deemed possible by Hankamer and Yang.

Once it is acknowledged that human memory capacity apparently is much larger than previously thought, Word and Paradigm morphology (Blevins, 2003, 2013, 2016; Matthews, 1974) turns out to provide a perspective on word formation in which the present experimental results can be readily integrated. Importantly, Word and Paradigm morphology is not a full listing theory, and we do not take our results to imply that lexical processing is driven primarily or exclusively by full listing. In fact, it may not be necessary to assume that all inflectional forms have representations in the brain much like entries in a lexical database. Work on naive discriminative learning has clarified that whole-word frequency effects as well as paradigmatic effects can arise as a consequence of language users learning to discriminate between experiences of the world on the basis of sublexical units (Arnold et al., 2017; Baayen et al., 2011, 2017).

A further question is, however, what are the consequences of the present experimental findings for morphological theory? The answer to this question depends on one's perspective on the role of morphological theory. If this role is conceived of as providing an insightful and succint explanation of the internal structure of morphologically complex words, then evidence from experiments on lexical processing is irrelevant. For instance, if lexical processing would proceed exclusively on the basis of full listing of all forms (which we do not think is the case), this would not provide any insights that would be useful for language education, for understanding the way language changes over time, or the evolutionary forces that have shaped modern languages.

All in all, what in our opinion the field of morphology is not served by is discussions of the experimental literature that are based on a strategy of discrediting and dismissing experimental evidence, as exemplified by Yang (2016)[p. 238]:

> "[. . .] the axiomatic and deductive nature of linguistics marks a clean break from the traditional methods in the social and behavioral sciences, which continue to loop through the cycle of data collection, statistical analysis, and repeat. In the best kind of linguistic practice, sim-

> ple hypotheses can be formulated precisely such that their empirical consequences of nontrivial depth can be worked out by mechanical means. Theoretical developments take place well before the collection and verification of data [. . .]. Occasionally, we do come across general principles of language that connect a wide range of phenomena; no need to bake each separately into the theory, or to invoke yet another variable in the model of regression.

Axiomatic and deductive theories may be useful and insightful, but this does not make them psychologically real. Importantly, they do not need to be psychologically real to be useful and insightful. However, when axiomatic and deductive theories are put forward as functional theories of cognitive processes, experimental evidence on cognitive processing is essential, and should not be dismissed when inconvenient. To those who find the post-Bloomfieldian construct of the morpheme not only attractive from the point of view of linguistic theory, but also attractive as a mental construct, the present empirical results will seem to be yet another frustrating example of the same old loop of "data collection, statistical analysis, and repeat" - frustrating not only because it contradicts empirical results supporting morpheme-based models (e.g., Marantz 2013), but also because they do not offer any insights into the questions that lie at the heart of generative theories of language. On the other hand, the present results fit well with non-decompositional theories according to which the forms of words arise in a system governed by the opposing communicative forces of predictability and discriminability (Blevins et al., 2017).

The main purpose of a morphological system is to serve its speakers. Hence, we think that when the goal is to understand language processing, linguistic, psycholinguistic as well as computational theories of morphology should inform each other regarding the most accurate principles of how morphologically complex words work.

Chapter 4

The time-course of Estonian morphological processing: insights from eye movements and pupillometry

Abstract

The present study investigated the time-course of Estonian morphological processing using eye gaze and pupil size measurements during a large-scale word naming task. We had two main goals: 1) to study when in the time-course of processing whole-word frequency and inflectional paradigm size effects begin to emerge in gaze and pupil dilation data, and 2) to explore the possibilities of using pupillometry for studying morphological processing more generally. The results of the study revealed that whole-word frequency emerged relatively early, as reflected in the second fixation duration, whereas inflectional paradigm size emerged later, as reflected in the total fixation duration. These findings are in line with Lõo, Järvikivi, Tomaschek, Tucker & Baayen (2017)'s quantile regression analysis of production latencies of the same experiment. They reported that whole-word frequency affects earlier processing, while inflectional paradigm size affects later processing. The findings from pupil dilation data showed similar results: both whole-word frequency and inflectional paradigm size effects emerged as significant predictors. However, inflectional paradigm size was a more reliable predictor of pupil dilation than whole-word frequency.

4.1 Introduction

There is a growing body of evidence indicating that not only is information about the morphological structure available in the mental lexicon, but also information about whole-word frequency and the paradigm relations of complex words. This has been shown for both languages with relatively simple morphology, such as English (Baayen et al., 2007) and Dutch (Baayen et al., 1997), and also for languages with more complex morphology, such as Serbian (Milin et al., 2009), Finnish (Moscoso del Prado Martín et al., 2004), and Estonian (Lõo et al., 2017).

For instance, Milin et al. (2009) studied Serbian inflectional paradigms with a lexical decision task and found that inflectional paradigms with higher inflectional entropy (i.e., forms occur roughly equally often within a paradigm) were recognised faster than inflectional paradigms with lower entropy. Moscoso del Prado Martín et al. (2004) investigated morphological families (e.g., *worker*, *workforce*, *handwork* etc) in Finnish and found that words with more family members were recognised faster than words with less family members. Finally, Lõo et al. (2017) examined the processing of Estonian case-inflected nouns, also using a lexical decision task. They found that higher whole-word frequency and larger inflectional paradigm size (i.e., the number of inflected forms in use for a given paradigm) decreased both response times and error rates. Additionally, inflectional paradigm size was negatively correlated with response latencies and error rates in a semantic categorization task. Lõo et al. (2017) replicated this finding with a word naming task, showing that inflected forms with higher whole-word frequency and larger inflectional paradigm size were produced with faster production latencies and shorter acoustic durations. However, these studies tell us relatively little about the time-course of processing, i.e., when certain effects become relevant during processing and, more specifically, how these effects emerge in time. Using quantile regression analysis of production latencies of the same experiment, Lõo et al. (2017) found that whole-word frequency was predictive of short production latencies, indicating that the effect was more relevant earlier in the processing, and that inflectional paradigm size was predictive of longer production latencies, indicating this effect was more relevant during later processing. The current paper investigates the time-course of Estonian case-inflected noun processing, by examining eye tracking data (fixation and pupillometry) collected during a large-scale word naming task (Experiment 2 in Lõo et al. 2017).

Until recently, much of the research on the time-course of morphological processing has used the masked priming paradigm, which is taken to reflect early stages of processing (see e.g., Beyersmann et al. 2016; Kazanina 2011; Lázaro et al. 2016; Longtin et al. 2003; Marslen-Wilson et al. 2008; Rastle & Davis 2008; Rastle et al. 2000, 2004; Stockall & Marantz 2006). Proponents of the early obligatory decomposition view (Rastle & Davis, 2008; Rastle et al., 2000, 2004) argue that all complex words must first undergo a form-based morphological decomposition. This is suggested not only for all regular complex words (e.g., *kill-er*), but also for pseudo-complex words (e.g., *corn-er*). In this approach, effects reflecting the whole complex word, including semantic transparency and whole-word frequency, can only be explained as late effects, possibly reflecting integration processes occurring at or after the recombination of morphemes (Fruchter & Marantz, 2015; Longtin & Meunier, 2005; Taft, 2004). Further evidence in support of decomposition comes from priming studies combined with electroencephalography (EEG) and magnetoencephalography (MEG) (Fruchter & Marantz, 2015; Lavric, Clapp & Rastle, 2007). They argue that orthographic and morphological sensitivity occurs earlier, at 100-200 ms post stimulus onset, and sensitivity to semantics and whole-word frequency arises later, at 300-500 ms post stimulus onset (but see Schmidtke et al. 2017 for possible shortcomings of these studies).

However, 300-500 ms post stimulus onset seems relatively late compared to studies reporting much earlier semantic effects for simple words. For instance, Kryuchkova, Tucker, Wurm & Baayen (2012) found that words with high danger ratings (e.g., *theft*) already showed positivity in the ERP signal at 150ms post stimulus onset. In another ERP study, using a semantic categorization task, Segalowitz & Zheng (2009) found an early lexicality effect at 100 ms and semantic access by 168 ms. Finally, early semantic effects have also been found for complex words. Davis, Libben & Segalowitz (2014) detected an effect of semantic transparency in compound processing around 100 ms in a lexical decision task combined with ERP.

As inflectional paradigm size is a novel measure, no precise claims about the time-course of this effect have been made, however, previous research suggests that it may be a later effect. First, morphological family size, a measure related to paradigm size, has usually been taken to be semantic in nature. This research suggests that the morphological family size effect is driven to a large extent by semantically related (semantically transparent) family members (Moscoso del Prado Martín et al., 2004; Mulder et al., 2014). Second, results from a word naming study by Lõo et al. (2017) showed that although effects of morphological family size and inflectional paradigm size were already present in the first quantiles, they were the strongest in the final quantiles. Finally, Schmidtke et al. (2017) conducted a distributional survival analysis of lexical decision latencies for English and Dutch derived words and found that whole-word frequency emerged the earliest (419 ms), followed by stem (i.e,. morphemic constituent) frequency (437 ms), and paradigmatic effects (491 ms).

The above results are in line with earlier research suggesting there are various semantic effects pertaining to the whole complex word forms, at various points during the course of processing (Feldman, 2000; Feldman, Kostić, Gvozdenović, OConnor & del Prado Martín, 2012; Feldman et al., 2015, 2009; Feldman & Soltano, 1999; Järvikivi & Pyykkönen, 2011; Järvikivi, Pyykkönen & Niemi, 2009; Rueckl & Aicher, 2008). Rueckl & Aicher (2008) found no effects in long lag priming for pseudo-complex words, unlike for transparent complex words. Järvikivi et al. (2009) reported that priming for pseudo-derived words was smaller than for derived words. In another Finnish masked priming study, Järvikivi & Pyykkönen (2011) found that pseudo-complex words resulted in significantly less priming than inflected words, and that there was an inhibition effect when the morphological family size of the pseudo-complex prime was larger than the morphological family size of the target. Further evidence challenging early decomposition is provided by Feldman and colleagues who showed that if one controls for the stem and affix properties between the different conditions then semantically transparent pairs show stronger priming than opaque pairs in English (Feldman et al., 2009);

the same was shown for Serbian (Feldman et al., 2012). In another study, Feldman et al. (2015) varied the stimulus onset asynchrony (SOA) between the target and the prime for semantically similar and dissimilar pairs, and showed that semantic priming increased linearly with increasing SOA (34–100 ms). Importantly, reaction times were already significantly faster for semantically related pairs at SOAs of 34 ms and 48 ms. Finally, a recent priming study by Milin, Feldman, Ramscar, Hendrix & Baayen (2017) found comparable priming effects for pseudo-derived words (e.g., *corn-er*) and orthographic controls (e.g., *brothel*). Taken together, these studies suggest that information about whole word form, and only about morphemic constituents, is relevant in online language processing, and possibly very early on.

Further evidence for early word form effects comes from eye tracking studies, Pollatsek, Hyönä & Bertram (2000) conducted a sentence reading study in Finnish, using 11-15 character long transparent noun-compounds, and found that the first fixation duration of Finnish compounds was affected by whole-word frequency as early as around 200 ms, and that the effect was not delayed compared to their monomorphemic matches. As indicated by another Finnish reading study, effects of whole-word frequency during the first fixation duration were further supported by using up to 18 character long compounds as stimuli (Kuperman, Bertram & Baayen, 2008). This is interesting because readers seemed to be sensitive to the properties of the whole word, long before they would have been able to read the whole word form. Furthermore, in a Dutch eye tracking experiment, Kuperman et al. (2009) observed whole-word frequency effects in a lexical decision task with isolated Dutch compounds. Finally, Juhasz & Berkowitz (2011) was one of the few studies that investigated the time-course of paradigmatic effects with eye tracking. In an English compound study, they showed that the left constituent family size of the compound influenced the likelihood of refixation and first fixation duration (see also Kuperman, Bertram & Baayen 2010).

4.1.1 The current study

The current study investigated the time course of Estonian case-inflected noun processing using eye tracking. Participants' eye movements and pupil size were recorded while they read aloud single inflected words.

Whereas behavioural measures, such as a production latency, provide a single predictor variable, gaze data provides multiple dependent variables (i.e., first fixation duration, second fixation duration, total fixation duration), allowing for further insight into the time-course of processing. Additionally, the current study makes use of a further eye tracking measure, pupil dilation. Instead of inspecting only pupil size peaks, like most of the previous research, we treated pupil size as a continuous variable and inspected pupil size changes over time. Since pupillometry is still relatively new and rarely used in psycholinguistic research, we will provide a brief overview of the previous relevant literature, before we will move on to the current study.

4.1.2 Pupillometry

A number of studies since the 1960s have shown that greater pupil dilation is correlated with higher cognitive load and emotional affect (see e.g., Ahern & Beatty 1979; Hess & Polt 1960; Kahnemann & Beatty 1966). More recently, researchers have started using pupil size measurement as an indicator of linguistic complexity (Engelhardt, Ferreira & Patsenko, 2010b; Hyönä et al., 1995; Just & Carpenter, 1993; Zekveld et al., 2010).

Just & Carpenter (1993) were among the first to use pupillometry in psycholinguistic research. They contrasted the processing of simpler with more complex sentences, and found that more complex sentences induced a larger pupil size. Hyönä et al. (1995) conducted a Finnish-English translation task and found that more difficult translations evoked a larger pupil dilation. Zekveld et al. (2010) investigated the relationship between pupil dilation and speech intelligibility, using a task where participants had to listen to sentences in noisy conditions of different levels. They found that the noisier the signal, the more difficult it was for the participant to make sense of the utterance, as indicated by stronger pupil responses (see Schmidtke 2017 for a review of linguistic studies using pupillometry).

A few studies have also used pupillometry in lexical processing research. For instance, Kuchinke et al. (2007) combined pupillometry with a lexical decision task, using high and low frequency words varying in emotional valence. Even though the response data showed an effect of both word frequency and valence, the pupil response was only affected by frequency. Lower frequency words induced a stronger pupil response, whereas valence did not affect the pupil response. Kuchinke et al. (2007) argued that the absence of valence effects in pupil dilation might due the fact that valence affects processing at a very early stage, such that is not sensitive to pupillary response.

Papesh & Goldinger (2012) detected frequency effects in the pupillary response during a delayed naming task. In this task, participants were asked to read aloud a target word presented on a screen, but they had to wait until a tone was played before making the response. The results of the pupillometry analysis indicated that pupil size increased across all trials, in response to the decreasing frequency of the target word for both speech planning and production. Geller et al. (2016) conducted a masked priming study with orthographically similar (e.g., *blur-blue*) and dissimilar (e.g., *blur-time*) prime-target pairs and found that pupil response was stronger for trials with orthographically similar pairs compared to dissimilar pairs and suggested that this reflected larger competition in similar compared to dissimilar pairs.

In summary, previous research on lexical processing has established that cognitive load associated with lower frequency and orthographically similar words induce a stronger pupil response. However, previous research has been conducted with simple words, and only investigated the timing and amplitude of a single pupil peak. However, because changes in pupil diameter, as a result of cognitive load, present a slow moving fluctuation which often peaks relatively late, it is not clear to which extent peak based measures offer insight into the time course of these processes.

Regarding the time-course of pupil dilation, previous research suggests that the pupil constricts in response to high luminance occur within 200 ms and that the response peaks at around 500 ms to 1,000 ms. Task-evoked pupillary responses are assumed to emerge from 200-300 ms, and reach their peak 500-1,000 ms post response (Beatty & Lucero-Wagoner, 2000). This indicates that pupil size is rather a slow measure researching its peak up to 1,000 ms after the stimulus onset. The exact timing of pupil peak seems to depend on the task. In the current study, we will investigate pupil size both across the whole trial window (0-4,000 ms) and across the time-course of the first 600 ms from the stimulus onset.

The present study is the first to investigate morphological processing using pupillometry. Specifically, we will examine whether lexical-distributional properties, such as whole-word frequency and inflectional paradigm size of the complex word, are reflected in the time course of pupil dilation.

4.2 Experiment

4.2.1 Participants

Thirty-three native speakers of Estonian (20 females; age 22-69, mean=38.34) with normal or corrected-to normal vision and no speech impairments participated in the study. Data for one participant was removed from the analysis, as he did not complete the experiment due to technical difficulties. Participants were recruited both from the University of Tallinn and Tallinn University of Technology, and they received 15 euros for their participation.

4.2.2 Materials

2,800 case-inflected nouns were selected from the Balanced Corpus of Estonian[1]. The whole-word frequency distribution of the items ranged between 1 and 1000 per million (median 40). The length of the stimuli varied between 2 and 19 characters (mean=7.88 characters). Stimuli were divided over 28 experimental lists, and each contained 400 items. Lemma frequency ranged between 1 and 3439 per million (median 519). Compounds were excluded from the data set. Inflected forms with both simplex and complex stems were added to the data set.

[1] http://www.cl.ut.ee/korpused/grammatikakorpus/ (15.04.2017)

4.2.3 Apparatus

The experiment was conducted in a sound attenuated booth. Participants' responses were recorded with a Marantz PMD670 digital recorder, using a supercarioid condenser table top microphone by Beyerdynamic, placed approximately 10 cm from participants' mouths. The eye movements were collected with an EyeLink II head-mounted eye tracker by SR Research, which is an infrared video-based tracking system. It consists of three miniature cameras mounted on a headband. Two cameras with infra-red LEDs are mounted on a headband to illuminate both eyes. The cameras measure pupil location and pupil size at a rate of 500 Hz. The registration is performed by placing the cameras and light sources 4-6 cm away from the eyes. Eye position data is accessed with 3.0 ms delay and the average spacial accuracy is approximately 0.5 degrees of arc. The third camera, a head-tracking camera is mounted on the centre of the headband at the level of the forehead. Four LEDs are attached to the corners of the computer screen to detect possible head movements and to automatically correct eye movement records. In the present study, only the right eye was tracked (except for 3 participants, where the left eye was the dominant). A nine-point calibration was used. The stimuli were presented on a 21-inch Dell computer screen in lower case 26-point Courier New Bold font using the ExperimentBuilder software by SR Research.

4.2.4 Procedure

Participants were instructed to read aloud as fast and as naturally as possible single words appearing on the computer screen. The naming data of the current experiment is reported in Lõo et al. (2017). Each trial started with a drift correction on the left of the screen, after which the target appeared in the center. The target stayed on the screen for 1,500 ms and was then replaced by a fixation cross, which remained on the screen for 2,500 ms to allow for the pupil size to resume to the baseline. The experiment started with ten practice trials, which were followed by 400 experimental trials. Every 100th trial was followed by a short break. The break lasted until the participant indicated that he was ready to continue. The whole procedure lasted approximately 90 minutes.

4.2.5 Analysis

Previous research using eye-tracking methodology in reading has established that first and second fixation duration can be used as indicators of earlier stages of reading, while total reading time and the number of fixations can be used as indicators later stages of a word reading (Bertram, 2011; Rayner, Chace, Slattery & Ashby, 2006; Rayner, Sereno, Morris, Schmauder & Clifton Jr, 1989).

In the current analysis, the following response variables were analysed: (1) first fixation duration (the duration of initial fixation in milliseconds); (2) second fixation duration (the duration of the second fixation in milliseconds); and (3) total fixation duration (the sum of fixation durations in milliseconds prior the end of articulation). The key predictors were: (1) whole-word frequency, which counted the total number of tokens for a particular word form, for example, the plural nominative *jalad* 'feet' occurred 1067 times, and the elative singular *jalast* 'from the foot' 149 times in the Balanced Corpus of Estonian; (2) lemma frequency, which is the summed frequency of all inflected words belonging to a particular inflectional paradigm, for example, the summed frequency of *jalg* 'foot/leg' is 5139; (3) Inflectional paradigm size, which is the number of inflected forms used for a certain lemma, for example, the inflectional paradigm size for *jalg* 'foot/leg' is 36. Additionally, orthographic length, participant's age, x- and y-coordinates of the fixations on the screen were added to the model as control variables.

Because task-independent differences in the pupil size measure were accounted for by adding by-subject random smooths for time, we decided to analyse the pupil dilation data without adding a baseline. However, adding the baseline (i.e., measuring the mean pupil dilation value between the drift correction and when the word appeared on the screen) to the model specification was also tested and it did not change the results.

First fixations below 50 ms and above 1,500 ms (3% of the data), second fixations below 50 ms and above 1,500 ms (1% of the data), and total fixations below 50 ms and above 2,000 ms (6.6 % of the data) were identified as outliers and removed from the analysis. Trials with misspoken stimuli, eye movement spikes and saccades (6.3% of the trials) due to eye-blinks, were removed from the pupil analysis. Finally, pupil measurements were down-sampled from every 2 ms (500 Hz) to every 20 ms (50 Hz) in order to reduce the amount of data. The final pupil model was checked for autocorrelation in the residual error.

To reduce the effects of extremely short and long fixation times, first and second fixation durations were transformed to roughly normal distribution using the logarithmic transformation. Whole-word frequency and lemma frequencies were also transformed using Johnson transformation (version 1.3, R package *Johnson*, Fernandez 2014).

Generalized additive mixed effects regression models (GAMMs) were used to analyse the data (Wood 2006, R-package *mgcv*, see Baayen et al. 2017). We opted for GAMM analysis, as it does not assume linearity between the predictor and response variables. This is particularly relevant for capturing time-course data such as the pupil dilation data. The output of GAMM consists of two parts: a parametric part and a non-parametric part. The parametric part is identical to linear models. The non-parametric part covers various non-linear phenomena, such as interactions, intercept adjustments, random smooths and random slopes. The size and direction of these effects can be best estimated through visualization.

First fixation duration

A GAMM fitted to the log-transformed first fixation duration data showed a U-shaped nonlinear effect of orthographic length. Short and long words had longer first fixation durations compared to medium length words. This effect levelled off at both ends (see Figure 4.1). Neither lemma frequency, whole-word frequency, nor inflectional paradigm size had a significant effect on the first fixation duration.

As a control variable, an interaction between x- and y-coordinates of the first fixation on the screen was added. The model also included by-participant factor smooths for trial, as well as by-participant random slopes for length. Further main and random effects (including by-item random intercepts) did not reach significance. The complete model summary can be found in Table 4.1.

Table 4.1: Summary of the partial effects in GAMM fitted to log-transformed first fixation duration.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.42	0.05	112.28	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Orthographic length)	4.39	5.37	6.33	< 0.0001
s(Orthograph length,Participant)	23.05	31.00	2.90	< 0.0001
s(X-coordinate, Y-coordinate)	23.29	27.06	63.68	< 0.0001
s(Trial,Participant)	163.96	287.00	230.18	< 0.0001

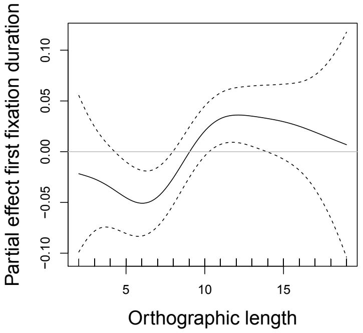

Figure 4.1: Partial effects for orthographic length in first fixation duration.The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

Second fixation duration

A GAMM fitted to the log-transformed second fixation duration showed a nonlinear effect of orthographic length. Longer words were read with a shorter second fixation duration. The left panel of Figure 4.2 shows that this effect levelled off when the words are very long. Further, a nonlinear whole-word frequency effect was detected. The middle panel of Figure 4.2 shows that more frequent words were read with shorter second fixation durations, however, this effect was not significant at the higher frequency range as indicated by very wide confidence intervals. Lemma frequency was also a significant predictor in the model. However, whole-word frequency outperformed lemma frequency (AIC-score difference was 2.60). Due to the high correlation between these two measures ($r = 0.58$), only whole-word frequency was included in the model. Furthermore, there was a significant interaction between x- and y-coordinates, and between participant age and orthographic length. The color coding in the right panel of Figure 4.2 indicates that younger participants' second fixation durations were longer compared to participants in the middle age range. The yellow colour marks longer fixation durations and blue shorter durations. For instance, for the orthographic length 5 on the x-axis, we can see that the colour changes from yellow to green and back to yellow as we move down the y-axis, indicating that both younger and older participants have longer second fixations with short words. Finally, the model included by-participant factor smooths for trial, as well as by-participant random slopes for length. Further main and random effects did not reach significance. The complete model summary can be found in Table 4.2.

Table 4.2: Summary of the partial effects in GAMM fitted to log-transformed second fixation duration

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.6361	0.0842	66.9324	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Orthographic length)	4.99	5.95	43.61	< 0.0001
s(Whole-word frequency)	1.58	1.94	7.48	0.002
s(Participant age)	1.00	1.00	0.30	0.58
ti(Participant age, Orthographic length)	9.71	11.66	2.62	0.002
s(X-coordinate, Y-coordinate)	15.17	19.92	35.43	< 0.0001
s(Orthographic length,Participant)	25.96	30.00	8.18	< 0.0001
s(Trial, Participant)	142.95	286.00	997.73	< 0.0001

Total fixation duration

Total fixation duration increased with orthographic length (see the left panel of Figure 4.3), and this effect was close to linear. The model was less confident with very long words as indicated by the wide confidence intervals. Total fixation duration decreased linearly with inflectional paradigm size (see the middle panel of Figure 4.3). Words with more paradigm member were read with shorter total duration. Neither whole-word frequency nor lemma frequency were significant predictors. Participant age and orthographic length entered into a nonlinear interaction. The

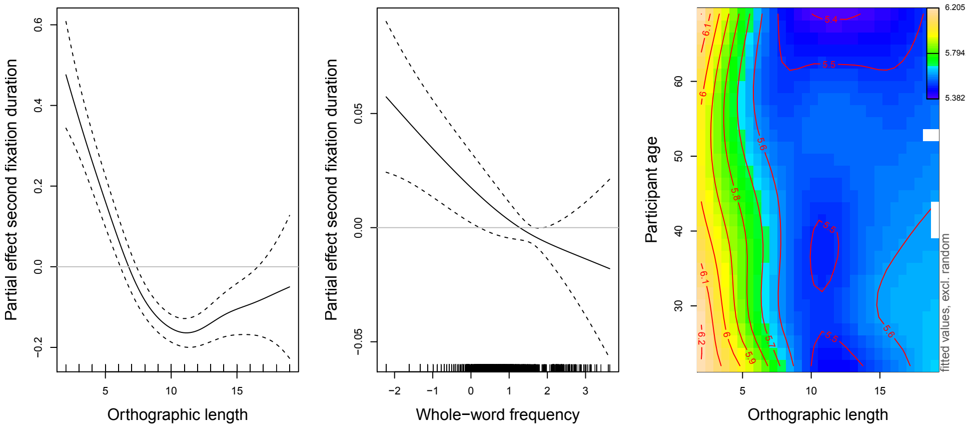

Figure 4.2: Partial effects for orthographic length, whole-word frequency and the tensor product between orthographic length and participant age in second fixation duration. The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

right panel of Figure 4.3 shows that older participants took less time reading short words, this is indicated the blue colour of semi-circle at the upper left side of the panel. Yellow indicates longer total fixation durations and blue shorter total fixation duration. The model also included by-participant factor smooths for trial, by-participant random slopes for length and inflectional paradigm size, as well as by-item random intercepts. Further main or random effects did not reach significance. The complete model summary can be found in Table 4.3.

Table 4.3: Summary of the partial effects in GAMM fitted to log-transformed total fixation duration

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1178.78	38.17	30.88	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Orthographic length)	3.99	4.75	33.27	< 0.0001
s(Inflectional paradigm size)	1.01	1.02	6.17	0.01
s(Participant age)	1.00	1.00	0.11	0.74
ti(Orthographic length,Participant age)	9.34	11.38	4.77	< 0.0001
s(Trial, Participant age)	200.83	286.00	564.47	< 0.0001
s(Orthographic length,Participant)	24.20	30.00	5.56	< 0.0001
s(Inflectional paradigm size,Participant)	18.68	31.00	1.58	< 0.0001
s(Item)	465.52	2734.00	0.21	< 0.0001

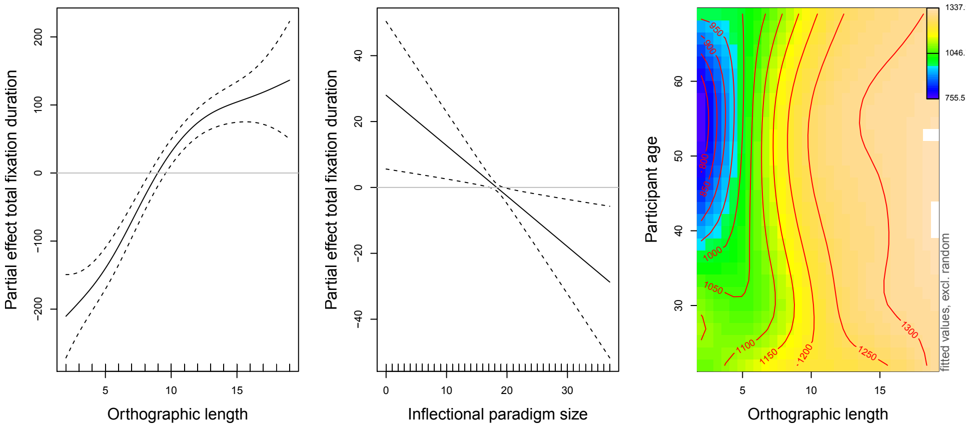

Figure 4.3: Partial effects for orthographic length, inflectional paradigm size and tensor product between orthographic length and participant age for total fixation duration. The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

Pupil dilation

A GAMM, fitted to the pupil dilation data comprising the whole 4,000 ms time-period of a trial, showed a range of fixed and random effects. First, there was a significant effect of time. The left panel of the upper row in Figure 4.4 shows that two pupil peaks emerged in the time-course of the trial: the first peak occurs around 300 ms and the second one around 1,500 ms. Second, there was a significant main effect of trial. The second left panel of upper row in Figure 4.4 indicates that pupil response decreased with trial, but the effect levelled off around the 100th trial. Pupil size decreased linearly with participant age, as indicated by the third left panel of the upper row in Figure 4.4, and it increased with orthographic length. The effect of orthographic length was close to linear. The effect was mostly triggered by medium length words and it levelled off at both ends (see the fourth left panel in Figure 4.4). Pupil size decreased nonlinearly with increasing inflectional paradigm size. The left panel of the middle row in Figure 4.4 shows that this effect was not present at the higher paradigm size range.

Furthermore, there was a significant nonlinear interaction between participant age and orthographic length. The yellow color in GAMM contour plots represents larger pupil size and blue smaller pupil size. Thus, the yellow triangle in the second left panel of the middle row indicates that whereas younger participants had overall a larger pupil size, this effect was the strongest for longer words. The model also indicated that there was a significant interaction between time and orthographic length, as well between time and inflectional paradigm size (the third and the fourth left panel of Figure 4.4). The relatively late peaks in both panels indicated that the effects reached their peaks relatively late. Words with longer orthographic length and smaller inflectional paradigm size increased pupil dilation. The model also included an interaction between x- and y-coordinates of eye gaze as controls. Finally, several random effects were included. Both by-time random smooths for participant (see the left panel of the lower row in Figure 4.4) and by-trial random smooths for participant (see the second left panel in the lower row in Figure 4.4) indicated that there were large differences how participants proceeded through trials and time. Finally, by-time random slopes for item were included (see the the third left panel of Figure 4.4). The complete model summary can be found in Table 4.4.

Table 4.4: Summary of partial effects in GAMM fitted to the pupil size during the time-period of 4,000 ms.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1229.40	48.66	25.26	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Time)	8.93	8.94	313.10	< 0.0001
s(Trial index)	8.67	8.70	1133.61	< 0.0001
s(Orthographic length)	8.91	8.99	84.34	< 0.0001
s(Participant age)	1.00	1.00	5.54	0.02
s(Inflectional paradigm size)	8.67	8.92	28.86	< 0.0001
ti(Participant age, Orthographic length)	15.71	15.98	143.19	< 0.0001
ti(Time, Orthographic length)	14.58	14.92	935.01	< 0.0001
ti(Time, Inflectional paradigm size)	14.26	14.80	81.99	< 0.0001
s(X-coordinate, Y-coordinate)	28.80	28.99	554.52	< 0.0001
s(Time, Participant)	255.75	278.00	1813.53	< 0.0001
s(Trial index, Participant)	257.43	277.00	157746.89	< 0.0001
s(Time, Item)	2699.53	2727.00	122.38	< 0.0001

In summary, a GAMM fitted to the complete time-series of trial indicated that there was a nonlinear main effect of inflectional paradigm size, but surprisingly

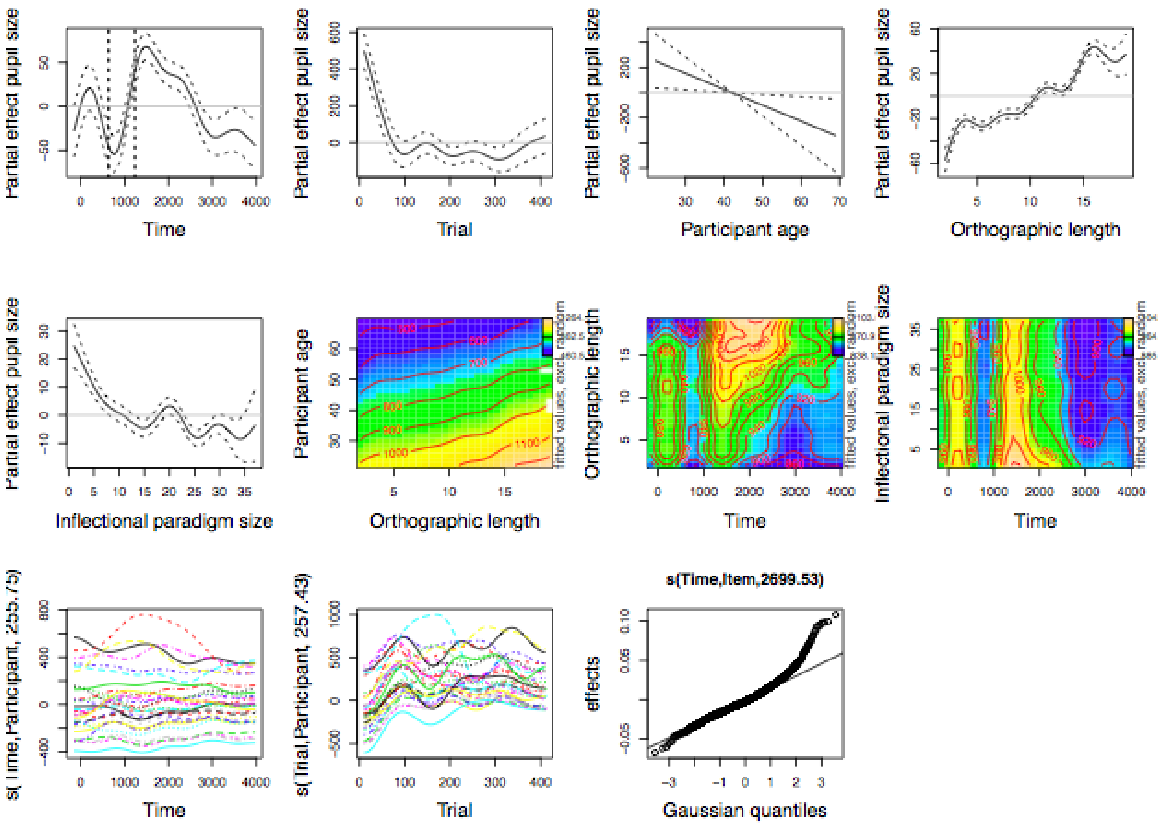

Figure 4.4: Partial effects for time (vertical horizontal lines represent speech onset and off set), trial, participant age, orthographic length, inflectional paradigm size and the tensor product between orthographic length and participant age, time and orthographic length and time and paradigm size; by-time random smooths for participant, by-trial random smooths for participant, and by-time random slopes for item.The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

no main effect of whole-word frequency. There was also no visible interaction between time and and whole-word frequency.

A possible reason for the lack of whole-word frequency effect is that the standard Gaussian GAMM might not be the most accurate statistical technique to make predictions about this complex data set. First, pupil size is not normally distributed, rather, the distribution is right-skewed with a long right tail. Additionally, we were not able to transform pupil size measure to normality using standard transformation techniques, such as log-transformation or *RE.Johnson*. Second, pupil data had many missing values due to blinks and spikes, in particular, very early and late in the time-course. This might have made it difficult for a GAMM model to make accurate predictors over the time-course of the whole trial (0 ms - 4,000 ms). Third, the residuals of the model did not follow a normal distribution and auto-correlation was still visible in the residuals, even after including the rho-parameter to the model. These observations led us to explore the data further using a Bayesian statistical technique, quantile regression analysis.

Quantile regression analysis of pupil size

Quantile regression (using with the R-package *qgam* by Fasiolo et al. 2016), is a distribution-free Bayesian technique. A quantile regression model does not assume any structure from the residuals or response variables. The current version of *qgam* is not yet able to handle large datasets. As the current data set included over two million data points even after down sampling, we therefore restricted our analysis to a fixed time-window. We selected the window of 0-600 ms after stimulus onset, as this is the time in which the naming response has usually been executed and the participant has already processed the word. This is indicated by the upper left panel of Figure 4.4, which shows that the first vertical dotted line is around 600 ms, indicating median speech onset.

A model fitted to the median quantile of the pupil dilation during the time window of the first 600 ms revealed a facilitatory nonlinear effect of whole-word frequency. The left panel of Figure 4.5 shows that this effect was mostly affecting the higher frequency range. Furthermore, there was also a facilitatory nonlinear effect of inflectional paradigm size. As the middle panel of Figure 4.5 indicates, the effect had the strongest effect at the smaller paradigm size range, but levelled off at a higher inflectional paradigm size range. Participant age had a weak linear effect on the pupil size (the right panel of Figure 4.5). Older participants had a smaller pupil size compared to younger participants. Finally, tensor product smooths for orthographic length and time, and by-participant random intercept were added. Further random structure (e.g., per-item intercepts or further time-smooths) was not added due to the limitations of the *qgam*-package. The complete model summary can be found in Table 4.5.

In summary, quantile regression analysis confirmed that the facilitatory effects

Table 4.5: Summary of partial effects in qgam fitted to the 0.5 decile of the pupil size during the time-period of 600 ms.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1165.5537	80.4822	14.4821	< 0.0001
B. smooth terms	edf	Ref.df	Chi.sq.-value	p-value
te(Time, Orthographic length)	16.89	19.63	1341.29	< 0.0001
s(Participant age)	1.12	1.12	6.39	0.01
s(Inflectional paradigm size)	2.69	2.93	34.52	< 0.0001
s(Whole-word frequency)	2.92	2.99	35.46	< 0.0001
s(Participant)	28.87	29.00	1840257.89	< 0.0001

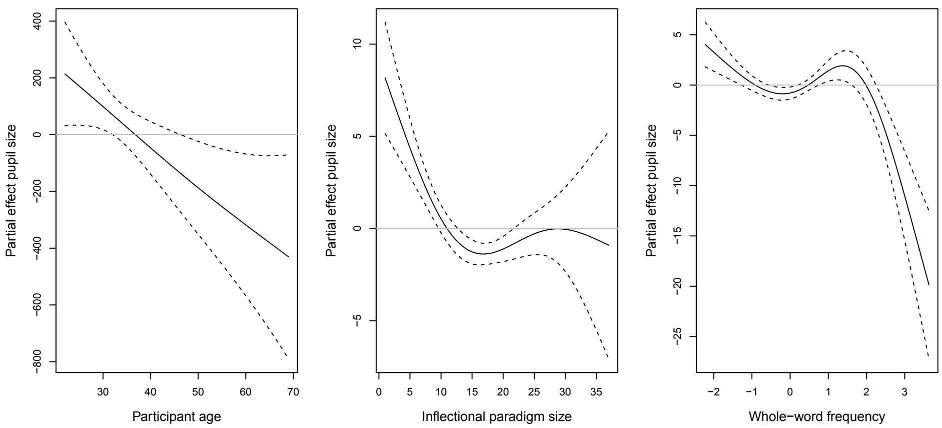

Figure 4.5: Partial effects for participant age, inflectional paradigm size and whole-word frequency in pupil size. The solid horizontal line represents the zero effect and dashed lines represent 95% confidence bands of the regression line for individual predictors.

of participant age and inflectional paradigm size were also present in the GAMM-analysis. The effect of whole-word frequency emerged in the quantile regression analysis, but not in GAMM-analysis.

4.3 General Discussion

The current study investigated the time-course of lexical processing in Estonian, with the main interest of investigating how whole-word frequency and inflectional paradigm size effects evolve over time.

Concerning the very early stages of reading, only an orthographic length effect emerged in very early stages of reading, as measured by first fixation duration.

Contrary to previous Finnish compound studies (Kuperman et al., 2008, 2009), we were not able to find very early effects of whole-word frequency for Estonian inflected forms. Interestingly, all lexical effects were missing in the first fixation duration data. The absence of early lexical effects was also supported with nonsignificant random intercepts for item. Additional eye tracking studies on inflected forms need to determine whether it is a side-effect of the experimental setup. Unlike in most previous studies, participants were reading words aloud and not silently. Further, before participants were presented with words in the middle of the screen, they needed to fixate on the left side of the screen, so, it took them some time to draw their focus back to the middle of the screen, which may also have resulted in less than optimal landing on the word and a quick refixation. Furthermore, words were displayed one by one without a context, whereas most previous eye tracking experiments have presented words in a sentential context. It is possible that early effects are not visible in the first fixation due the nature of reading aloud task. For instance, it might take participants longer to adjust to reading aloud task compared reading silently, which causes delays in lexical effects.

In the second fixation duration data, lexical effects arose in the form of the whole-word frequency effect. Also lemma frequency arose as a significant predictor of second fixation duration, however, whole-word frequency was a better predictor than lemma frequency (AIC was 2.6 lower). In the total fixation duration data, we established an effect of inflectional paradigm size, words with larger inflectional paradigm size were read with a shorter total fixation duration. Interestingly, neither lemma frequency nor whole-word frequency reached significance in total fixation analysis. This temporal locus of whole-word frequency and inflectional paradigm size fits well with findings from word naming analyses of the same experiment in Lõo et al. (2017), which showed that whole-word frequency had a stronger effect for shorter latencies whereas inflectional paradigm size had a stronger effect for longer responses.

In summary, our fixation data analyses were unable to find support for early form-driven decomposition (see e.g., Beyersmann et al. 2016; Longtin et al. 2003; Rastle et al. 2004). Instead, we observed an early whole-word frequency effect, which according to the decompositional account could only emerge due to late integration of morphemic components (see e.g., Fruchter & Marantz 2015 and Taft 2004). With the additional effect of paradigm size found in total fixation durations, our findings are in line with a growing line of research reporting early semantic effects for complex words (Davis et al., 2014; Feldman, 2000; Feldman et al., 2015, 2009; Järvikivi & Pyykkönen, 2011). Moreover, early effects of whole-word frequency in our fixation data fit well with other studies reporting whole-word frequency effects in early fixations of compound words (Kuperman et al., 2008, 2009), as well as with work by Lõo et al. (2017) and Schmidtke et al. (2017) who found that whole-word frequency emerged earlier and paradigmatic effects later in response time analyses of word naming and lexical decision task.

Pupil size analysis was in most part in line with the eye movement data. In overall pupil size analysis, the only significant predictor was inflectional paradigm size. However, both whole-word frequency and inflectional paradigm size effects were detected when the analysis window was restricted to 0-600 ms after stimulus onset. Contrary to what we expected, there was no interaction between time and whole-word frequency. This might be due to the fact that fixation data is more precise compared to pupil dilation data where effects seem to emerge with a time-lag. Another reason might be individual differences in how these effects evolve over time for individual participants.

Using per-subject and per-group pupil data analyses, Lõo, van Rij, Järvikivi & Baayen (2016) found individual differences with respect to frequency effects. Some participants showed a strong sensitivity to the frequency of the stimulus, whereas for others the effect was almost absent. They suggested that individual differences in lexical processing may arise due to differences in participants reading strategies. The current study suggests that another individual difference measure that seems to affect lexical processing is participant age. Age effects in Estonian lexical processing already emerged in a lexical decision task by Lõo et al. (2017), which showed that older participants were slower, but more accurate when recognising Estonian case-inflected forms. Using a found individual differences between individual pupil dilation patterns, these differences were associated with differences in lexical processing.

Our results showed that older participants had weaker pupillary responses while reading aloud words compared to younger participants. Additionally, there was an interaction between orthographic length and participant age in both eye movement and pupil size analyses. Older participants had a shorter total duration when reading short words compared to younger participants. At the same time, younger participants showed a stronger pupil size effect, indicating higher processing load, when they read longer words. This could be possibly be explained by the fact that older participants have more experience with language (Ramscar et al., 2014, 2017). Older participants in our experiment have seen more words than younger participants, which might results in their improved ability to read larger word units with less effort (see also Spieler & Balota 2000).

To our knowledge, the current study was the first one to investigate morphological processing using pupillometry. Therefore, a few issues regarding the use of pupillometry in morphological processing research should be also pointed out. First of all, pupillometry is a slow measure Beatty & Lucero-Wagoner (2000). Thus, it may not be the most suitable method for studying the time course of morphological processing, as there is a relatively long lag between the time when a response takes place, and when it becomes visible in the pupil data. For the exact time-course of morphological processing, fixation data seems to be more informative. However, since pupillometry is a continuous measure, similar to EEG, tracking pupil size differences over time can inform us about the relative time-course of processing. Pupil measurements are, however, also less labor intensive and time-consuming to collect, thus it can potentially serve as an alternative to to EEG when looking at the time-course of lexical processing. Second, measuring pupil size during word naming made it difficult to tease apart the effects of the task and the effects of language processing. Other tasks, such as sentence reading and lexical decision with morphologically complex words, should be conducted in combination with pupillometry to eliminate the possibility that effects found in the current study were not due to lexical processing. Third, regarding statistical analysis of pupil data, more research is certainty needed to determine how to analyse the abnormal, highly variable and complex nature of this data. For instance, more research is needed to determine possible advantages of using both GAMM-s and quantile regression on analysing pupil data.

In summary, the current eye tracking study revealed that early fixations were modulated by whole-word frequency, and later fixations by inflectional paradigm size. Pupil size did not show sensitivity to whole-word frequency in GAMM analysis, but it did in quantile regression. The analysis of pupil size during the first 600 ms using quantile regression suggests that both effects of frequency and paradigm size affect pupil responses. In particular, pupil dilation, but also gaze data analyses suggest that there are also large individual differences in the processing patterns of individual speakers. These results are not easily reconcilable in models that assume early morpheme-based obligatory decomposition for inflected words, instead they suggest that both whole-word constitution and word meaning affect processing of inflected forms early on. Moreover, in line with other recent approaches reporting similar results, we find that a word's morphological relatives have a robust effect on word recognition and production.

Chapter 5

Individual differences in pupil dilation during naming task

Abstract

The present study investigates individual differences in pupil dilation during standard word naming. We looked at (i) how individual subjects' pupil size changes over the course of time and (ii) how well pupil size is predicted by the frequency of the stimuli. The time course of the pupil size was analysed with generalized additive modeling. The results show large individual variations in the pupil response pattern in this very simple task. Although, we see a pupil response to both stimulus onset and articulation onset and offset, both the amplitude of change and the direction of change differ substantially between subjects. This raises the question of what makes the pupil response functions so diverse, and one factor indicated by the frequency effect or the lack thereof might be shallow reading versus reading for content.

5.1 Introduction

The eye's pupil diameter size reflects changes in luminance, emotional state, and also cognitive processes (Ahern & Beatty, 1979; Hess & Polt, 1960; Kahnemann & Beatty, 1966; Young & Biersdorf, 1954). For example, Kahnemann & Beatty (1966) conducted a recall task to show that pupil size can be used to measure mental effort. The recall of more complex number strings triggered a stronger pupil response compared to the recall of less complex strings. Pupil dilation has also been found to reflect *linguistic* processing, such as the complexity of a linguistic task (Hyönä et al., 1995), sentence intelligibility (Zekveld et al., 2010), sentence complexity (Ben-Nun, 1986; Just & Carpenter, 1993; Schluroff et al., 1986) and spoken language comprehension (Engelhardt et al., 2010a).

Finally, pupil diameter has been used to investigate lexical processing costs (Geller et al., 2016; Kuchinke et al., 2007; Papesh & Goldinger, 2012). Kuchinke et al. (2007) measured pupil dilation in high and low frequency words and found that high frequency words receive higher peak pupil dilation compared to low frequency words. Papesh & Goldinger (2012) replicated the effect, using delayed naming paradigm, in which they delayed the time of the naming response up to 2000 ms and observed that frequency affects pupil size, even after the naming responses were issued. Geller et al. (2016) reported for a masked priming study that higher-frequency words elicited an earlier dilation of the pupil compared to low-frequency words. They also observed that dilation increased when words had more lexical competitors, for both low- and high-frequency words.

The present study investigates changes in pupil size while participants read out loud words presented to them one by one on a computer screen, using an inter-stimuli interval of around 4000 ms in order to allow the pupil to contract to baseline after each trial. The present study addresses the question of the extent to which the response to cognitive load in lexical processing varies across individual participants. The response of the pupil to a cognitive task has been described mathematically by Hoeks & Levelt (1993) as single smoothly increasing and then slowly decreasing function of time. Wierda, van Rijn, Taatgen & Martens (2012) argued that a time series of pupil dilation values can be a superposition of several such functions, arising as a consequence of several cognitive events taking place within the window of time under investigation.

If in a controlled task, such as word naming, the pupil indeed responds in a fixed and consistent way to the demands of the task (reading the word, preparing for articulation, controlling the articulation itself, and preparation for the next trial), then one would expect the same pupil dilation function for all participants, possibly with minor variations in shape, similar to minor variations in intercept and slope that one may detect for linear response functions with the help of (generalized) linear mixed models. However, if reading styles differ substantially across participants, a general response function may not be appropriate. In what follows, we address these questions using generalized additive mixed modeling (Wood 2006; see also Baayen, van Rij, de Cat & Wood 2016). First, however, we introduce the details of our naming experiment.

5.2 Word naming experiment

5.2.1 Participants

Thirty-three native speakers of Estonian (18 women; 22-69 years; mean age 38), with normal or corrected to normal vision and no diagnosed speech impairments, participated in the experiment. The participant's dominant eye was tracked (30 right eye, 3 left eye). Participants received 15 euros for their participation.

5.2.2 Materials and design

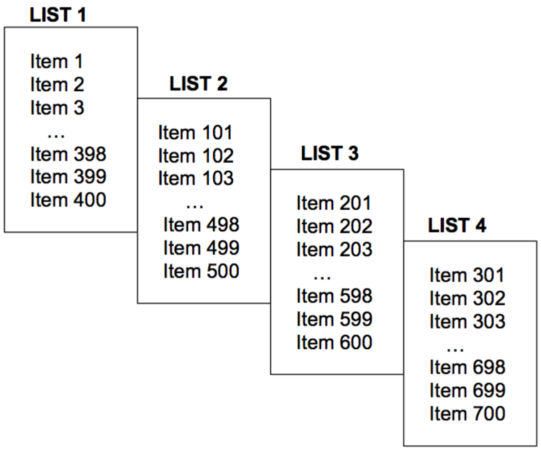

Figure 5.1: Outline of the experimental design.

A total of 2800 case-inflected nouns were randomly selected from the Balanced Corpus of Estonian. The frequency distribution resembled the distribution of the complete corpus, and ranged between 1 and 1000 per million (mean 14.27). The length of stimuli varied between 2 and 19 characters (mean 7.88; sd 2.62 characters). Twenty-eight experimental lists were created from the randomized set of 2800 items, each with 400 words. To maximize the number of items in the experiment, an overlapping design was used (see Figure 5.1), with a 300-word overlap between successive lists. A given word occurred four times in the experiment, once in each of the four lists.

5.2.3 Apparatus

The experiment was conducted in a medium illuminated sound-attenuated booth. Eye movements and pupil size were recorded using the head-mounted EyeLink II eye tracker by SR Research Ltd. EyeLink II is a video-based tracking system with a resolution of 500 Hz, ca 3.0 ms delay, and an average spacial accuracy of approximately 0.5 degrees of arc.

The naming data was recorded separately from the stimulus presentation program ExperimentBuilder by SR Research with a Marantz PMD670 digital recorder, using a supercarioid condenser table top microphone by Beyerdynamic, placed approximately 10 cm from the participant's mouth.

5.2.4 Procedure

Participants were tested individually. First, they were familiarized with the procedure: reading aloud, as naturally as possible with words presented one by one on the computer screen. They were asked to start speaking as soon as the word appeared on the screen. As participants were wearing a head-mounted eye tracker, they were instructed to move their head as little as possible.

Participants were seated in front of the computer screen at a distance of approximately 60 cm. The experimenter placed the headband of the eye tracker over participant's head and adjusted it such that the eye position could be correctly tracked on the computer screen. Further, the eye tracking system was calibrated. Adjustments were made until the spacial accuracy of the eye location measurement was smaller than 0.5 degrees of arc. The stimuli were presented on a 21-inch Dell grey background computer screen in black lower case 26-point Courier New

Bold font. The screen resolution was 1024x768 pixels.

Each trial started with a drift correction on the left of the screen, after which the target appeared in the center. The target stayed on the screen for 1500 ms and was then replaced by a fixation cross that remained on the screen for 2500 ms. Thus, in total each trial lasted 4000 ms. We extended the length of each trial to ensure that the pupil size was recorded with enough time delay for the pupil to contract to the baseline. The experiment started with ten practice trials, which were followed by the 400 experimental trials. Every 100th trial was followed by a short break. At the end of the experiment participants filled out a language background questionnaire. The whole procedure lasted approximately 90 minutes.

5.2.5 Data preparation and analysis

Naming latency and articulation duration were calculated directly from the audio recordings using Matlab (version 8.5.0). Prior to the analysis, data for two participants were removed due to technical problems during tracking. Thus, the analysis was conducted on the data from 31 subjects. Trials with misspoken stimuli, eye movement spikes and saccades due to eye-blinks were removed from the analysis using R (version 3.2.1, 6.3% of the trials). The statistical analysis was performed with R (version 3.2.1, using the *mgcv* package, version 1.8.6 of 2015, for generalized additive mixed regression (GAMM) modeling Wood (2006); see also Baayen et al. (2016), for visualization, we made use of the *itsadug* package.

The dependent variable of interest was log-transformed *Pupil size*, measured in the standard (arbitrary) units delivered by the eye tracking system. The main predictors were *Time* in milliseconds and *Frequency*. Frequency was transformed to normality using the Johnson transformation (version 1.3, R package *Johnson*, Fernandez 2014). *Gaze coordinates* (x- and y-axis position on the screen in pixels) were added to account for changes in measured pupil size due to the location on the screen.

For each participant, we fitted a separate GAMM to the 400 time series of pupil dilation values (resulting in 31 models). In addition to a general smooth for a subject's pupil dilation curve, we also included, for each word, a nonlinear random effect curve in time, using shrunk factor smooths. The X and Y coordinates of the fixation position were included as controls using a tensor product smooth, and as we anticipated the effect of frequency to vary over time, we also included a tensor product smooth for frequency and time. The model was checked for autocorrelation in the residual error, and an AR (1) autocorrelation parameter was then added to the model to remove, as far as possible, autocorrelative structure from the residuals (see Baayen et al. (2016) for a detailed discussion).

In what follows, we first discuss the estimated pupil dilation functions for the ensemble of 31 subjects. We then provide more details on three subject groups resulting from a clustering analysis.

5.2.6 Individual patterns over time

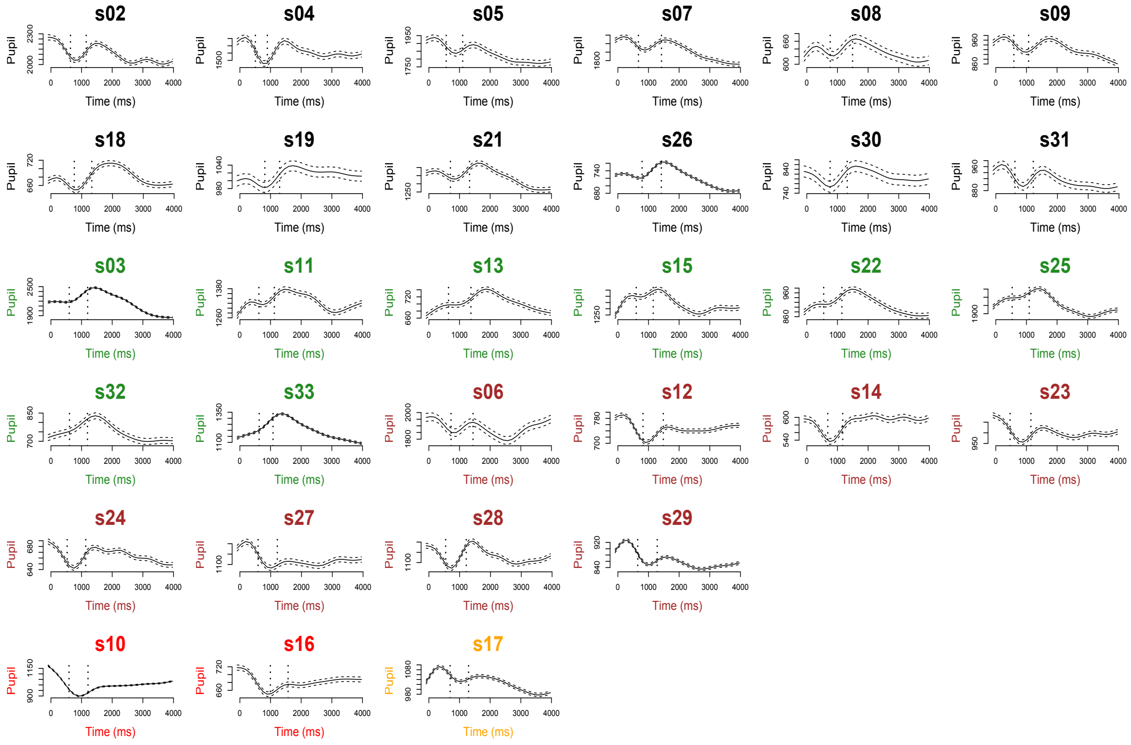

Figure 5.2: The fitted effects of time smooths for 31 subjects, five groups are indicated by different color-coding. The first vertical dotted line indicates the median articulation onset and the second dotted line the median articulation offset.

Figure 5.2 presents the estimated pupil dilation functions provided by GAMMs that focused on the main effect of time (leaving out frequency as covariate). The x-axis represents time and the y-axis presents pupil size. The stimulus was presented at time 0 ms, and a trial ended after 4000 ms. The first black dotted vertical line in a panel represents the median onset of articulation and the second dotted line the median offset of articulation.

Figure 5.2 shows that the relation between the Pupil size and the Time differs substantially between subjects. The different dilation functions fall into five groups, obtained with a divisive hierarchical clustering method using Manhattan distance applied to the first three principal components of a principal component analysis of the correlation matrix of the empirical first derivative of the subject time smooths, and indicated by different color coding.

Group 1. The first group is the largest (12 subjects: s02, s04, s05, s07, s08, s09, s18, s19, s21, s26, s30 and s31). For these subjects, the pupil dilation function shows a first peak shortly after stimulus onset and a second peak at or shortly after the onset of articulation. However, some subjects show a slightly different pattern from this general trend. For subject s02, s21, s26 and s30 there is no clear first peak, and for subject s04 and s09 the second peak much later after the onset of articulation.

Group 2. The second group includes eight subjects (s03, s11, s13, s15, s22, s25, s32, s33). Here only one clear pupil response is visible, which occurs slightly after speech onset and peaks after speech offset. Somewhat different are subject s11 and s15 who also show a slight peak after stimulus onset. However, the main difference is a well-differentiated initial peak in the pupil dilation function which is present in group 1, but absent in group 2.

Group 3. The third group also includes eight subjects (s06, s12, s14, s23, s24, s27, s28 and s29). Compared to the first two groups, this group is more heterogeneous. Also here, only one pupil peak is visible and it occurs after speech onset. However, unlike in the second group, in this group pupil size is declining instead of increasing from stimulus onset. Some subjects again deviate from the general group pattern. For example, for subject s14 there is no peak after speech onset, but the pupil stays at high plateau even after the articulation offset. Furthermore, pupil fluctuation patterns of subject s27 and s29 are somewhat similar to the first group as also here two clear pupil responses are present. However, for these subjects the second peak is much smaller than the second peak in the first group. Finally, unlike the first two groups, the relative heights of the pupil maxima are quite diverse.

Other dilation curves. The last two groups obtained by the hierarchical clustering technique clearly stand out from the rest (see the last row in Figure 5.2: subject s10, s16 and s17). Subject s10 and s16 start with a high pupil size, but it declines significantly after the stimulus onset. The pupil size increases also only a little after the articulation onset and stays relatively constant without a clear second peak. Finally, the last cluster only included subject s17. This subject seems to be similar to the first group and also has an initial pupil peak. However, like subject s10 and s16, this subject has no clear second peak. Because the two last groups include only three subjects, we excluded them from further group analysis.

In the present experiment, there are three key events to which the pupil appears to be sensitive: the presentation of the stimulus, the onset of articulation, and the offset of articulation. The first three groups show an increase in pupil size around the onset of articulation. The first group also shows an increase in pupil size after the word is presented, whereas it is less so for the rest of the four groups. The difference between the second and third is group is that although for both a second peak is present, subjects in the second group have an increasing pupil size and subjects in the third group decreasing pupil size. Finally, subject s10, s16 and s17 (group 4 and 5) start with a decline in pupil size from stimulus onset and show only a small increase at articulation onset.

The subject-specific pupil dilation functions indicate that even though participants engage in exactly the same task, with exactly the same procedure, they apparently engage in this task in cognitively significantly different ways. Subjects who show little or no dilation following stimulus onset may be experienced readers, and subjects who show a decreasing pupil dilation function may be highly skilled talkers.

Next, we included a smooth for frequency and a tensor product smooth for frequency and time to the GAMM model. We compared the frequency effect differences between the three largest subject groups.

5.2.7 Frequency effects in subject groups

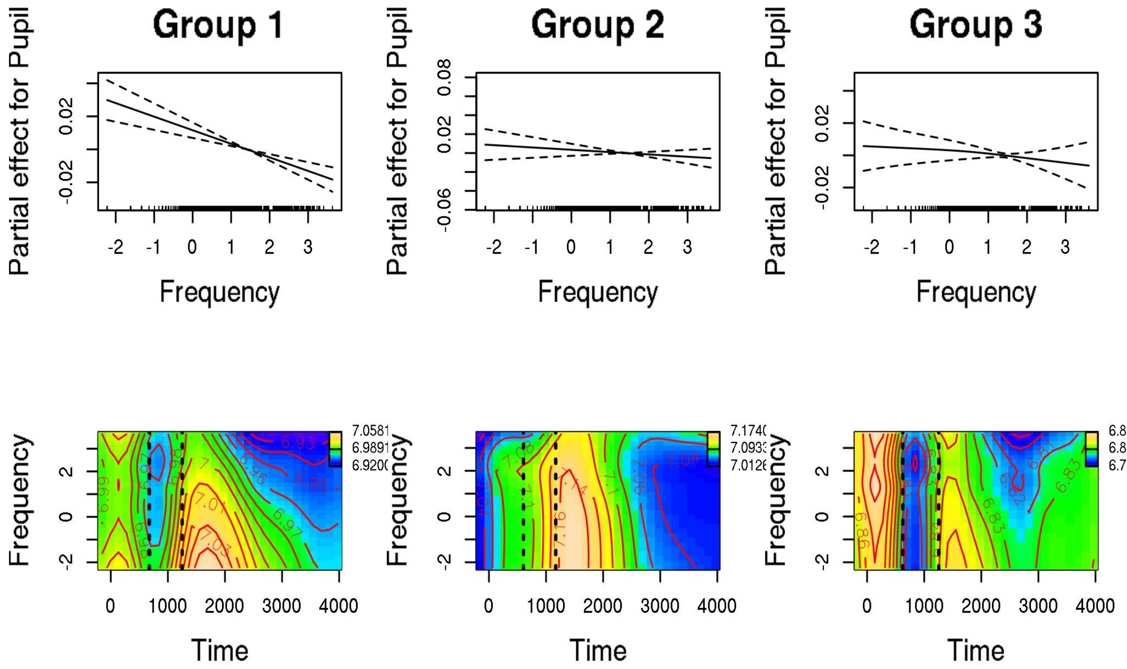

Figure 5.3: The partial effects of Frequency (the upper panels) and the tensor product smooth for Frequency and Time for three subject groups without random effects. The first vertical dotted line is the median articulation onset, the second vertical dotted line is the median articulation offset.

Figure 5.3 presents the main partial effect of frequency (top panels) and the way frequency over time modifies the pupil dilation function for the three main subject groups (bottom panels). As in Figure 5.2, the first vertical dotted line represents median articulation onset and the second line median articulation offset. The frequency measure, normalized (and thus centered), ranges from -2 to +3; pupil size was log-transformed.

Group 1. The summary of the statistical model for the first group indicated a significant main effect of Frequency (t(898846)=-4.93; p-value < 0.0001). Pupil dilation decreases linearly with increasing frequency, which is consistent with high-frequency words being cognitively less demanding, typically affording shorter responses latencies in the word naming task (Forster & Chambers, 1973).

Frequency entered into a significant nonlinear interaction with time (F(14.02, 896425.4)=8.72; p-value $<$ 0.0001). The way in which frequency modulates the pupil dilation is presented in the bottom-left panel of Figure 5.3. The contour plot can be read like a topographic map with peaks and valleys, yellow indicates the highest and blue the lowest elevation. The contour lines show the slope of the Pupil size as a function of Time and Frequency. The lines that are closely spaced represent steep slope and contour lines further apart gentler slopes, i.e., slower changes in Pupil size.

When the value on the y-axis is kept constant at zero, we see a similar pattern to the simple time smooths for members of the first group in Figure 5.2. The pupil starts off with a slight peak in pupil size shortly after the stimulus onset, contracts before the articulation, dilates again during articulation and finally, contracts at the end of the trial. This is color-coded by changes from green to yellow, from yellow to green, from green back to yellow and from yellow back to green and then blue. However, as we can also see, the pupil dilation changes differently over time, dependent on the frequency of the word. The pupil size increases earlier in time and more with lower frequency words than higher frequency words (see the peak around 0 ms and the yellow peak around 1500 ms in the bottom-left panel of Figure 5.3).

It is noteworthy that the effect of frequency on pupil size is the largest for this group *after* speech offset. This can be seen by considering the gradient before and after articulation. Before articulation, we find three contour lines, after articulation, we find five. Even around 3000 ms after stimulus onset, frequency still has a strong effect. This suggests the effect of frequency is not restricted to the early stages of information uptake, and is likely to have a strong semantic component.

Group 2. As the middle panels of Figure 5.3 show, the second group has no main effect of Frequency (F(1, 578743.7)=1.17; p-value=0.28. However, the frequency and time interaction is significant (F(15.034, 578743.7)= 9.90; p-value $<$ 0.0001). The contour lines in the bottom-middle panel are all fairly vertical, indicating an effect of time and hardly any modulation by frequency, except a weak effect before speech onset at high frequency range.

This pattern of results suggests that the second group might not be semantically engaged. It is well known that experienced readers can read out text while thinking about other issues. We suspect that the members of the second group are rather 'mechanically' performing the task, but are not deeply engaged in interpretation, possibly because isolated words in a word naming experiment are out of context and have no communicative value.

Group 3. This group also shows no significant main effect of Frequency (*F*(1.255, 591182.6)=8.95; *p*-value=0.42), but a non-linear interaction (*F*(14.388, 591182.6)=7.52.; *p*-value <0.0001). The interaction is presented in the bottom-left panel of Figure 5.3.

The interaction effect of the third group is somewhat similar to the first group. However, compared to the first group, the effect is more shallow and gradient in particular during and after articulation. At stimulus onset, there is no frequency effect, but there is a gradual decline in pupil size across frequency span until the speech onset. Further, compared to the first group, we see fewer contour lines and a more gradual change from yellow to green after the end of the articulation. Also for this group the pupil dilation peak is weaker after speech offset. The results suggests more engagement with words compared to the second but less compared to the first group.

5.3 General Discussion

The results of this study can be summarized as follows. Inspection of the pupil dilation function for 31 participants revealed substantial variation in not only the magnitude of dilation in response to stimulus onset, speech onset, and articulation offset, but also in the direction of change, with some subjects showing contraction and others dilation. In the light of such substantial inter-subject variability, it makes little sense to try to extract a 'population dilation curve' from this kind of data. Such a curve would not come close to characterizing the pupil response function for many of the participants. This seems to apply in particular to pupillometry data, but also cognitive research in general (see e.g., Roehm, Bornkessel-

Schlesewsky, Rösler & Schlesewsky 2007, for similar conclusions in EEG data). What this shows is either that different subjects engage in exactly the same task in very different ways or that subjects are all engaged in exactly the same way, but their engagement is manifested differently in their pupil dilations. However, based on the group differences in frequency, we argue for the first option.

Given substantial variability in subjects' reading abilities (see e.g., Kuperman & Dyke 2011), loquaciousness, amount of education, social status and responsibilities, as well as differences in age, gender, and motivation for participating in a psycholinguistic experiment, these differences are perhaps unsurprising. But these differences clearly indicate that general statements about loci of processing effects in the 'population' based on pupillometry data are potentially hazardous, if the present pattern of results, revealed by detailed investigation with generalized additive mixed models, turn out to be replicable in future experiments.

This conclusion is supported by an examination of the frequency effect of three subject groups. Two of these subject groups (group 1 and 3) showed a frequency effect that was the strongest after the offset of articulation. One of the groups (group 1) also showed a weak frequency effect immediately after stimulus onset. The second group showed only a weak effect of frequency, even though the members of this group, just as the others, read the words and produced them correctly. The weak frequency effect, in combination with a relatively shallow pupil response, may be indicative of short but semantically shallow lexical processing (e.g., Baayen & Milin (2010), who observed the absence of a word frequency effect for fast readers, and the strongest effect of frequency for the slowest readers in self-paced reading of continuous text).

We think that for a proper understanding of lexical processing, in all its currently bewildering variability, it will be essential to consider in much more detail the vast differences in experience, motivation, socio-cultural background, as well as differences in brain morphology, that subjects bring into an experiment.

Chapter 6

General Discussion and Conclusions

The goal of the present dissertation was to investigate how native speakers of Estonian process case-inflected nouns in their native language. Estonian is a morphologically rich Finno-Ugric language with extremely productive morphology and an inflectional system that increases the number of possible word forms in Estonian to millions. In actual language use, however, not all forms are used equally, such that which case a particular noun takes, depends on its semantics. With these observations in mind, we set out to examine which lexical-semantic properties co-determine Estonian lexical processing. In particular, we were interested in the role whole-word frequency and inflectional paradigm size have in the processing of case inflected nouns.

The four studies provided converging evidence that whole-word frequency and inflectional paradigm size are the main predictors of Estonian case-inflected noun processing. In particular, evidence from pupillometry suggested that there are individual differences in how native speakers of Estonian process case-inflected nouns. In what follows, I will first summarize the results from the individual studies. Next, I will discuss the main research questions raised at the beginning of the current dissertation. Finally, I will suggest directions for future research and acknowledge the limitations of our studies, before providing a conclusion of the dissertation.

6.1 Summary of results

Study 1 investigated the comprehension of Estonian case-inflected forms, using a lexical decision (3,000 items; 24 participants, age 24-67 years) and a semantic categorization task (200 items; 26 participants, age 21-67 years), in order to investigate which lexical properties influence the response times and accuracies. The results of the lexical decision task indicated that forms with higher whole-word frequency and larger inflectional paradigm size were recognized faster and with fewer errors. In the semantic category task, forms with larger inflectional paradigm size were recognized faster and with fewer errors. Additionally, older participants were slower but more accurate in both tasks.

Study 2 examined the production of Estonian case-inflected forms with two word naming experiments. Experiment 1 (200 items; 26 participants, age 21-67 years) showed that forms with a larger morphological family size elicited shorter production latencies and forms with higher whole-word frequency elicited shorter acoustic durations. Experiment 2 (2,800 items; 33 participants, age 22-69 years) revealed that higher whole-word frequency, larger inflectional paradigm size and morphological family size decreased both production latencies and acoustic durations. Finally, a quantile regression analysis on production latencies indicated that whole-word frequency had a stronger effect earlier in the time-course of production, whereas inflectional paradigm size and morphological family size had stronger effects later in the time-course of production.

Study 3 investigated the time-course of Estonian morphological processing using eye movement and pupillometry measurements during word naming. The results indicated that during second fixation duration, whole-word frequency emerged as a significant predictor. The facilitatory effect of inflectional paradigm size emerged later in the time-course, as measured by total fixation duration. Pupillometry confirmed the facilitatory effect of inflectional paradigm size. In addition, both eye movement and pupil dilation analyses revealed interesting interaction effects between orthographic length and participant age. The analysis showed that older participants were faster in reading both very short and long words.

Study 4 explored the role of individual differences in morphological processing using both per-participant and per-group analyses of pupil dilation during word naming. First, the study examined how pupil dilation changes over time and found that pupil patterns differed substantially between participants. Second, using a hierarchical clustering technique, three groups of individual pupil patterns were established. In all groups the effect of whole-frequency was established. However, whereas some participant groups showed a strong frequency effect, others were only slightly affected by whole-word frequency.

6.2 Discussion of research questions

What kind of information is used by Estonian native speakers when they read and produce inflected forms?

The traditional psycholinguistic approaches to morphologically complex words differ in terms of how much and what information they assume is stored and what information is computed online. 1) Sublexical theories posit automatic decomposition (Beyersmann et al., 2016; Kazanina, 2011; Lázaro et al., 2016; Longtin et al., 2003; Marslen-Wilson et al., 2008; Rastle & Davis, 2008; Rastle et al., 2000, 2004) or across-the-board decomposition for all complex words (Fruchter & Marantz, 2015; Marantz, 2013; Solomyak & Marantz, 2010). 2) Supralexical accounts argue that all complex forms are first accessed by their whole-forms and then later decomposed into morphemic constituents (Giraudo & Grainger, 2000, 2001). 3) Dual-mechanism theories argue that complex words are either decomposed or stored depending on how regular or frequent they are (Clahsen, 1999; Clahsen et al., 2003; Hahne et al., 2006; Marcus et al., 1995; Niemi et al., 1994; Pinker, 1999; Ullman, 2001). 4) Dual-route models support both decomposition and storage, and the correct processing route is determined by various lexical distributional properties (Frauenfelder & Schreuder, 1992; Schreuder & Baayen, 1995).

Contrary to what most of these traditional approaches would predict, throughout our studies, the processing of Estonian case-inflected nouns was predicted by whole-word frequency and inflectional paradigm size. In the behavioural data, whole-word frequency was a significant predictor of lexical decision response times and accuracies, as well as word naming latencies and durations. In the eye tracking data, whole-word frequency emerged relatively early in the time-course, i.e., during second fixation duration. One one hand, this is different from what has usually been found for Finnish, another Finno-Ugric language closely related to Estonian. For instance, in Finnish, whole-word frequency effects have been standardly reported for compounds and derived forms (Bertram & Hyönä, 2003; Bertram et al., 1999; Järvikivi, Bertram & Niemi, 2006; Pollatsek et al., 2000; Vannest et al., 2002), but only for very high frequency inflected forms (Laine et al., 1995, 1999; Niemi et al., 1994; Soveri et al., 2007). However, these previous studies on Finnish have used factorial designs with maximally dozens of words. Thus, it is an open-ended question whether our findings are due to the larger scale design of our studies or due to the cross-linguistic differences between these two related languages.

On the other hand, our findings are very much in line with whole-word frequency effect found for regular inflected forms in other languages (Baayen et al., 1997; Balling & Baayen, 2008; Caselli et al., 2016). Furthermore, they also fit well with evidence showing that English multi-word sequences (e.g., *the president of*) are also subject to frequency effects (Arnon & Snider, 2010; Bannard & Matthews, 2008; Janssen & Barber, 2012; Tremblay et al., 2011; Tremblay & Tucker, 2011). This is particularly relevant, because English multi-word prepositional phrases (e.g., *into the house*) can often be translated into Estonian with a single word (e.g., *majasse*).

Importantly, we also showed that the number of attested paradigm members for a given word affected processing. We established a facilitatory effect of inflectional paradigm size in lexical decision and semantic categorization tasks, as well as in word naming. Evidence from eye tracking revealed that total fixation durations and pupil size decreased with increasing inflectional paradigm size. These finding fit well with other paradigmatic effects established by earlier research, such that morphological family size and inflectional entropy affect the processing of complex words across languages, word types and experimental paradigms (De Jong et al., 2002; Milin et al., 2009; Moscoso del Prado Martín et al., 2004;

Moscoso del Prado Martín et al., 2004; Schreuder & Baayen, 1997; Tabak et al., 2010). Neither decompositional nor storage-based accounts would be able to straightforwardly predict these effects. Especially since there seems to be no principled way of explaining how an obligatory decompositional account could easily account for these effects

However, at the same time, it also poses a similar question about the nature of mental representation, as finding whole-word frequency effects in Estonian and multi-word sequences in English. That is, when a rule-driven processing were in place, we would not expect to find these effects. Furthermore, storing all possible English multi-word sequences seems even more implausible than to storing all Estonian inflected forms, due to limited memory capacity. Thus, one needs to either acknowledge that humans are capable of representing vastly more information about words in long term memory compared to what has been deemed feasible, or search for alternatives to strict adherence to rules and representations.

When during the time-course does this information become available? Traditional decompositional models would not expect whole-word frequency to emerge at all, or at least not early on (Beyersmann et al., 2016; Fruchter & Marantz, 2015; Kazanina, 2011; Longtin et al., 2003; Marantz, 2013; Rastle & Davis, 2008; Rastle et al., 2004; Solomyak & Marantz, 2010; Taft & Forster, 1975, 1976b) Similarly, decompositional processing models would not expect paradigmatic effects, including inflectional paradigm size to arise. These accounts do not specify a mechanism for paradigms between complete whole-words.

Contrary to these assumptions, we established that whole-word frequency is the strongest predictor of earlier processing in both behavioural and eye tracking data, predicting both shorter production latencies of quantile regression and second fixation durations of gaze data. Our results fit with well with findings by Schmidtke et al. (2017), who reported that whole-word frequency emerged early on in a survival analysis with Dutch and English derived words. Further, our findings are also compatible with eye movement studies, such as Pollatsek et al. (2000), who reported early whole-word frequency effects for Finnish compounds and with a growing line of research reporting early semantic effects for complex words (Davis et al., 2014; Feldman, 2000; Feldman et al., 2015, 2009; Järvikivi & Pyykkönen, 2011).

Further, inflectional paradigm size and morphological family size were predictive of longer production latencies, total fixation durations and pupil dilation, indicating it emerges rather late in the time-course. These effect provide further evidence for the semantic nature such of paradigmatic measures, like morphological family size (De Jong, 2002; Moscoso del Prado Martín et al., 2004).

Do these lexical effects manifest themselves the same way for all participants?

Two kinds of evidence from the current dissertation indicate that lexical processing effects do not manifest themselves the same way for all people. First, throughout the dissertation, we were able to establish processing differences between younger and older participants. For example, we observed an interesting speed-accuracy trade-off in lexical decision and semantic categorization in Study 1. Older participants responded slower, but at the same time, they were also more accurate in both lexical decision and semantic categorization. This might be a reflection of their greater experience with language (Ramscar et al., 2014, 2017), rather than a sign of cognitive decline, as has been generally assumed. The slowing down for older participants might come with certain advantages, such as bigger vocabulary or better lexical knowledge.

The idea of potential cognitive advantage is further supported by our eye tracking data; although older participants were slower in responding to lexical decision and semantic categorization, their processing load during word naming was actually lower compared to younger participants. This has been shown by reduced pupil dilation for older participants compared to younger participants, in particular, older participants had an advantage in reading longer words, suggested by both gaze and pupil data.

Second, using individual subject and group analyses, Lõo et al. (2016) found individual differences with respect to how pupil size and word frequency interact with each other. Some participants had a strong facilitatory effect of frequency, However, for others the effect was almost absent, meaning pupil size did not reflect changes in the frequency of the stimulus. They suggested that these individual differences in lexical processing may arise due to differences in participants' reading strategies. However, future research needs to establish whether these differences might be due to differences in reading abilities, as suggested by Andrews & Lo (2013); Falkauskas & Kuperman (2015); Schmidtke et al. (2017).

6.3 Limitations

All studies in the current dissertation provide complementary and converging evidence about the processing of Estonian case-inflected nouns. However, a few potential limitations should be also pointed out.

First, our studies used isolated single words. This was done primarily to control for context effects, but also to make it easier to compare results across methodologies and across the bulk of previous studies across the past four decades. This decision, however, also had specific consequences. For instance, it made our studies less natural, as language speakers usually encounter words in speech or in text. In particular inflected forms (e.g., *jalgadel* 'on the feet') may sound odd without a context and they make more sense in a sentence such as in *Villid olid* ***jalgadel*** 'Blisters were **on the feet**'). Moreover, it makes a big difference *what* or *who* is *on the feet*, for example, *blisters* can be *on the feet*, but at the same time, someone can also be *on steady feet*, meaning *financially secured*. Some previous studies on Finnish have suggested that inflected words in isolation seem to be decomposed into morphemic components (Laine et al., 1995, 1999; Niemi et al., 1994; Soveri et al., 2007), however, other studies suggest this might not be the case with inflected forms in context. For instance, Hyönä, Vainio & Laine (2002) conducted a reading study with Finnish inflected forms presented in sentences, and showed that lexical decision and fixation times were similar between monomorphemic and polymorphemic Finnish inflected forms (see also Bertram, Hyönä & Laine 2000; Juhasz 2012; Luke & Christianson 2011). This suggests that inflected forms in a context are processed as whole forms, which is very much in line with whole-word frequency effects for inflected forms found in the current dissertation.

Further, our participants were more diverse compared to the typical participant pool of psycholinguistic experiments, which have predominantly tested undergraduate university students. All our participants varied greatly in age (21–69 years) and in various other aspects, such as language background and occupation. This made our findings on Estonian inflectional processing more generalizable to the general population of Estonian speakers. However, this also made it also more difficult to study certain aspects of language processing. For example, we suggested based on the previous recent research (Andrews & Lo, 2013; Falkauskas & Kuperman, 2015; Schmidtke et al., 2017) that individual differences in pupil patterns might be due to differences in reading skills. Although we collected information on participants language skills via self-reports in a language questionnaire, we could not associate this information with differences in reading patterns. More objective measures, such as vocabulary or reading proficiency are needed to test that hypothesis.

6.4 Future directions

First, all our studies examined morphological processing with visual word recognition and production. It would be interesting to examine the extent to which these results would generalize to auditory processing. In reading, the processing of inflected forms is more influenced by factors that are not available in auditory processing. For example, visual tasks, such as visual lexical decision or reading always include an additional occulomotor component resulting that information intake is different. Further, orthographic length, transparency and frequency have shown to influence visual processing of morphologically complex words more that auditory processing. In listening, information about larger chunks of the complex word becomes available incrementally as the speech signal unfolds over time, whereas, at the same time, it is possible to recognize a written word with a single fixation. Therefore, it would be interesting to see whether whole-word frequency and paradigmatic effects would emerge stronger or weaker in auditory word recognition.

Second, previous research has indicated that participants' rating scores on emotional valence (how positive a word is) and arousal (how exciting a word is) can co-determine lexical processing. For instance, Kuperman, Estes, Brysbaert & Warriner (2014) showed that more positive, negative and more exciting words facilitated lexical decision response times. However, to my knowledge, previous research in this area has been only conducted with Indo-European languages. It would be interesting to collect such ratings also from native speakers of other language families, such as Finno-Ugric. Furthermore, not many studies have looked at the emotional affect in combination with complex words. To my knowledge, only one study, Kuperman (2013) studied how valence and arousal affect the processing of English compounds. He found that the valence of both constituents, as well as the valence of the whole-word facilitated lexical decision times whether affective properties of co-determine processing for inflected words and for morphologically rich languages is yet to be determined.

Third, it would be important to conduct large-scale studies, similar to the present ones in other languages with rich inflection, for example, in other two large Finno-Ugric languages, Finnish or Hungarian. Both Finnish and Hungarian are closely related to Estonian, but have even more inflectional cases, 15 and 18 respectively. Therefore, it would be particularly interesting to examine whether the effect of inflectional paradigm size would also replicate in Finnish and Hungarian. Compared to Estonian, Finnish has more agglutinative forms, including a higher number of clitics and possessive markers. Therefore, finding full forms frequency effects in Finnish and Hungarian would be even more challenging for decompositional models compared to Estonian. As mentioned above, previous research on Finnish has been usually not able to find whole-word frequency effects for inflected forms (Laine et al., 1995, 1999; Niemi et al., 1994; Soveri et al., 2007). However, this might have been a statistical power issue (see e.g., Westfall et al. 2014) and these effects might emerge in studies with more items.

Fourth, recent research suggests that individual differences may play an important role in lexical processing. With the exception of a couple of studies (Andrews & Lo, 2013; Falkauskas & Kuperman, 2015; Schmidtke et al., 2017), this aspect has been rarely explored for morphologically complex words. In the current dissertation, we established individual differences with respect to frequency effect, whereas some participants showed a strong frequency effect, it was absent for others. Future research is needed to establish whether similar differences also emerge in other lexical distributional measures, such as whole-word frequency, inflectional paradigm size, and morphological family size. Further, research is needed to determine what causes individual differences in the processing of morphologically complex words. For instance, whether they might be caused by differences in reading strategies or reading skill like we suggested in the General Discussion of Study 4. Finally, other methods such as self-paced reading, where a context is also provided, may be needed to assess frequency effects.

Fifth, to gain more insights into the precise processing principles of Estonian case-inflected forms, simulating Estonian processing with computational models, such as the Naive Discriminative Learning model or other connectionist models might be informative. For instance, Naive Discriminative Learner successfully simulated morphological family size, inflectional entropy and various other effects in English, Dutch and Serbian (Baayen et al., 2011; Mulder et al., 2014). As a next step, it would be interesting to see whether it can also handle a morphologically rich language such as Estonian and predict for instance, the effects of whole-word frequency and inflectional paradigm size.

This brings us also to the question about the general relationship between linguistic theory, psycholinguistics and computational modeling of complex words. That is, how could psycholinguistics benefit more from morphological theory and computational modeling. The goal of linguistic theory is to describe the principles of how complex words work, and the goal of psycholinguistics and computational linguistics of complex words is to inquire into these and further principles experimentally and computationally. That is, whether these principles actually work in mind, and whether they generalizable to large amounts of actual language data. Most insight is gained when these three fields would work together.

6.5 Conclusions

Although morphological processing research has a long history, the field is currently more than ever open to new explorations. The current dissertation contributed to the field by studying the role of whole-words and inflectional paradigms in the processing of Estonian inflection. We were able to show that even a simple linguistic unit such a single word can be a rich source of information into the complexities of a language. However, further explorations using different languages and experimental paradigms are needed, as we are only beginning to uncover the precise underpinnings of all the linguistic and non-linguistic factors that contribute to the language processing mechanisms and the mental lexicon.

Bibliography

Ahern, S. & Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, *205*(4412), 1289–1292.

Alegre, M. & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, *40*, 41–61.

Allen, M. & Badecker, W. (1999). Stem homograph inhibition and stem allomorphy: Representing and processing inflected forms in a multilevel lexical system. *Journal of Memory and Language*, *41*(1), 105–123.

Allen, P. A., Madden, D. J., & Crozier, L. C. (1991). Adult age differences in letter-level and word-level processing. *Psychology and Aging*, *6*(2), 261.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(02), 239–273.

Anderson, S. R. (1977). On the formal description of inflection. In *Chicago Linguistic Society*, volume 13, (pp. 15–44).

Anderson, S. R. (1982). Where's morphology? *Linguistic inquiry*, *13*(4), 571–612.

Anderson, S. R. (1992). *A-morphous morphology*. Cambridge: Cambridge University Press.

Andrews, S. & Lo, S. (2013). Is morphological priming stronger for transparent than opaque words? it depends on individual differences in spelling and vocabulary. *Journal of Memory and Language*, *68*(3), 279–296.

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

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

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

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

Aronoff, M. (1992). *Morphology Now*. SUNY series in linguistics. State University of New York Press.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, Mass.: The MIT Press.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge, U.K.: Cambridge University Press.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*, 295–328.

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

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*, *36*, 94–117.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech*, *56*, 329–347.

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding*. Amsterdam/Philadelphia: Benjamins.

Baayen, R. H., Levelt, W., Schreuder, R., & Ernestus, M. (2008). Paradigmatic structure in speech production. *Proceedings Chicago Linguistics Society 43*, *1*, 1–29.

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

Baayen, R. H. & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Baayen, R. H., Milin, P., Filipovic Durdjevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481.

Baayen, R. H., Schreuder, R., De Jong, N. H., & Krott, A. (2002). Dutch inflection: the rules that prove the exception. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), *Storage and Computation in the Language Faculty* (pp. 61–92). Dordrecht: Kluwer Academic Publishers.

Baayen, R. H., Sering, T., Shaoul, C., & Milin, P. (2017). Language comprehension as a multiple label classification problem. In *Proceedings of the 32nd International Workshop on Statistical Modelling (IWSM)*, Johann Bernoulli Institute, Rijksuniversiteit Groningen, The Netherlands, 3–7 July 2017.

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

Baayen, R. H., Vasishth, S., Bates, D., & Kliegl, R. (2017). The cave of shadows. addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *56*, 206–234.

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*, 419–463.

Balling, L. & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*, *23*, 1159–1190.

Balling, L. W. & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, *133*, 283–316.

Balota, D. A. & Chumbley, J. I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, *24*, 89–106.

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

Beatty, J. & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of psychophysiology*, *2*, 142–162.

Ben-Nun, Y. (1986). The use of pupillometry in the study of on-line verbal processing: Evidence for depths of processing. *Brain and Language*, *28*(1), 1–11.

Bertram, R. (2011). Eye movements and morphological processing in reading. *The Mental Lexicon*, *6*(1).

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*, 390–405.

Bertram, R. & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language*, *615-634*, 48.

Bertram, R., Hyönä, J., & Laine, M. (2000). The role of context in morphological processing: Evidence from Finnish. *Language and Cognitive Processes*, *15*(4-5), 367–388.

Bertram, R., Laine, M., Baayen, R. H., Schreuder, R., & Hyönä, J. (1999). Affixal homonymy triggers full-form storage even with inflected words, even in a morphologically rich language. *Cognition*, *74*, B13–B25.

Beyersmann, E., Ziegler, J. C., Castles, A., Coltheart, M., Kezilas, Y., & Grainger, J. (2016). Morpho-orthographic segmentation without semantics. *Psychonomic Bulletin & Review*, *23*(2), 533–539.

Bien, H., Baayen, R. H., & Levelt, W. J. (2011). Frequency effects in the production of dutch deverbal adjectives and inflected verbs. *Language and Cognitive Processes*, *26*(4-6), 683–715.

Bien, H., Levelt, W., & Baayen, R. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences of the USA*, *102*, 17876–17881.

Blevins, J. P. (2003). Stems and paradigms. *Language*, *79*, 737–767.

Blevins, J. P. (2005). Word-based declensions in Estonian. In G. E. Booij & J. v. Marle (Eds.), *Yearbook of Morphology 2005* (pp. 1–25). Dordrecht: Springer.

Blevins, J. P. (2006). English inflection and derivation. In B. Aarts & A. M. McMahon (Eds.), *Handbook of English Linguistics* (pp. 507–536). London: Blackwell.

Blevins, J. P. (2008). Declension classes in estonian. Linguistica Uralica. Estonian Academy Publishers.

Blevins, J. P. (2013). Word-based morphology from Aristotle to modern WP (word and paradigm models). In K. Allan (Ed.), *The Oxford handbook of the history of linguistics* (pp. 375–395). Oxford: Oxfort University Press.

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

Blevins, J. P., Milin, P., & Ramscar, M. (2017). The zipfian paradigm cell filling problem. In J. Kiefer, J. P. Blevins, & H. Bartos (Eds.), *Perspectives on Morphological Organization: Data and Analyses* chapter 8. Leiden:Brill.

Bobaljik, J. D. (2012). *Universals in comparative morphology: Suppletion, superlatives, and the structure of words*. MIT Press.

Burani, C. & Caramazza, A. (1987). Representation and processing of derived words. Report 25, Cognitive Neuropsychology Laboratory, the Johns Hopkins University.

Burke, D. M. & Shafto, M. A. (2008). Language and aging. *The handbook of aging and cognition*, *3*, 373–443.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*(3), 297–332.

Caselli, N. K., Caselli, M. K., & Cohen-Goldberg, A. M. (2016). Inflected words in production: Evidence for a morphologically rich lexicon. *The Quarterly Journal of Experimental Psychology*, *69*(3), 432–454.

Chialant, D. & Caramazza, A. (1995). Where is morphology and how is it processed? the case of written word recognition. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 55–78). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Christianson, K., Johnson, R., & Rayner, K. (2005). Letter transpositions within and across morphemes. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(6), 1327–1339.

Clahsen, H. (1999). Lexical entries and rules of language: a multi-disciplinary study of German inflection. *Behavioral and Brain Sciences*, *22*, 991–1060.

Clahsen, H., Sonnenstuhl, I., & Blevins, J. P. (2003). Derivational morphology in the german mental lexicon: A dual mechanism account. *Morphological structure in language processing*, *151*, 125.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., & Mans, H. (2017). Conll-sigmorphon 2017 shared task: Universal morphological reinflection in 52 languages. *arXiv preprint arXiv:1706.09031*.

Dabrowska, E. (2004). Rules or schemas? evidence from Polish. *Language and cognitive processes*, *19*, 225–271.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2007). TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.1. Technical Report ILK 07-07, Computational Linguistics Tilburg University.

Davis, C. P., Libben, G., & Segalowitz, S. J. (2014). Compounding matters: The p1 as an index of semantic access to compound words. In *Psychophysiology*, volume 51, (pp. S34–S34).

De Jong, N. H. (2002). *Morphological families in the mental lexicon*. MPI Series in Psycholinguistics. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.

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*, 555–567.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*, 329–365.

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). Berlin: Mouton de Gruyter.

Di Sciullo, A.-M. & Williams, E. (1987). *On the definition of word*, volume 14. Springer.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2007). Do transposed-letter similarity effects occur at a morpheme level? evidence for morpho-orthographic decomposition. *Cognition*, *105*(3), 691–703.

Embick, D. & Noyer, R. (2007). Distributed morphology and the syntax/morphology interface. *The Oxford handbook of linguistic interfaces*, 289–324.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010a). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, *63*(4), 639–645.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010b). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, *63*(4), 639–645.

Erelt, M. (2003). *Estonian language*. Tallinn: Estonian Academy Publishers.

Erelt, M., Erelt, T., & Ross, K. (2007). *Eesti keele käsiraamat*. Eesti Keele Sihtasutus.

Erelt, M., Erelt, T., Viks, Ü., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., & Vare, S. (1995). *Eesti keele grammatika. 1., Morfoloogia sõnamoodustus*. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Erelt, T., Leemets, T., Mäearu, S., & Raadik, M. (2013). *Eesti õigekeelsussõnaraamat: ÕS 2013*. Eesti Keele Sihtasutus.

Falkauskas, K. & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing english compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1607.

Fasiolo, M., Y., G., Nedellec, R., & Wood, S. N. (2016). *Fast calibrated additive quantile regression*. R package version 1.0.

Feldman, L., Pastizzo, M., Soltano, E., & Francis, S. Semantic transparency influences morphological processing. *Brain and Language*, *in press*, xx–yy.

Feldman, L. B. (2000). Are morphological effects distinguishable from the effects of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1431–1444.

Feldman, L. B., Kostić, A., Gvozdenović, V., OConnor, P. A., & del Prado Martín, F. M. (2012). Semantic similarity influences early morphological priming in serbian: A challenge to form-then-meaning accounts of word recognition. *Psychonomic bulletin & review*, *19*(4), 668–676.

Feldman, L. B., Milin, P., Cho, K. W., del Prado Martín, F. M., & OConnor, P. A. (2015). Must analysis of meaning follow analysis of form? a time course analysis. *Frontiers in human neuroscience*, *9*.

Feldman, L. B., O'Connor, P. A., & Moscoso del Prado Martin, F. (2009). Early morphological processing is morpho-semantic and not simply morpho-orthographic: evidence from the masked priming paradigm. *Psychonomic Bulletin & Review*, *16*(4), 684–691.

Feldman, L. B. & Soltano, E. G. (1999). Morphological priming: The role of prime duration, semantic transparency and affix position. *Brain and Language*, *68*(1-2), 33–39.

Fernandez, E. S. (2014). *Harrell Miscellaneous*. R package version 1.4.

Forster, K. & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Fox, J. & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage.

Frauenfelder, U. H. & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. E. Booij & J. v. Marle (Eds.), *Yearbook of Morphology 1991* (pp. 165–183). Dordrecht: Kluwer Academic Publishers.

Fruchter, J. & Marantz, A. (2015). Decomposition, lookup, and recombination: Meg evidence for the full decomposition model of complex visual word recognition. *Brain and Language*, *143*, 81–96.

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

Geller, J., Still, M. L., & Morris, A. L. (2016). Eyes wide open: Pupil size as a proxy for inhibition in the masked-priming paradigm. *Memory & cognition*, *44*(4), 554–564.

Giraudo, H. & Grainger, J. (2000). Effects of prime word frequency and cumulative root frequency in masked morphological priming. *Language and Cognitive Processes*, *15*, 421–444.

Giraudo, H. & Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. *Psychonomic Bulletin and Review*, *8*, 127–131.

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.

Hahne, A., Mueller, J. L., & Clahsen, H. (2006). Morphological processing in a second language: Behavioral and event-related brain potential evidence for storage and decomposition. *Journal of cognitive neuroscience*, *18*(1), 121–134.

Halle, M. & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, volume 24 of *Current Studies in Linguistics* (pp. 111–176). Cambridge, Mass: MIT Press.

Hanique, I. & Ernestus, M. (2012). The role of morphology in acoustic reduction. *Lingue e Linguaggio*, *11*(2), 147–164.

Hankamer, J. (1989). Morphological parsing and the lexicon. In Marslen-Wilson, W. (Ed.), *Lexical Representation and Process*, (pp. 392–408)., Cambridge, MA, USA. MIT Press.

Harley, H. & Noyer, R. (1999). Distributed morphology. *Glot international*, *4*(4), 3–9.

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

Hay, J. B. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*, 1041–1070.

Hendrix, P. (2015). *Experimental Explorations of a Discrimination Learning Approach to Language Processing*. PhD thesis, University of Tbingen.

Hess, E. H. & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*(3423), 349–350.

Hockett, C. (1954). Two models of grammatical description. *Word*, *10*, 210–231.

Hoeks, B. & Levelt, W. J. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, *25*(1), 16–26.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. (2006). Survival ensembles. *Biostatistics*, *7*, 355–373.

Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, *48*(3), 598–612.

Hyönä, J., Vainio, S., & Laine, M. (2002). A morhological effect obtains for isolated words but not for words in sentence context. *European Journal of Cognitive Psychology*, *14*, 417–433.

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

Janssen, N., Bi, Y., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes*, *23*(7-8), 1191–1223.

Järvikivi, J., Bertram, R., & Niemi, J. (2006). Affixal salience and the processing of derivational morphology: The role of suffix allomorphy. *Language and cognitive processes*, *21*(4), 394–431.

Järvikivi, J. & Pyykkönen, P. (2011). Sub-and supralexical information in early phases of lexical access. *Frontiers in Psychology*, *2*.

Järvikivi, J., Pyykkönen, P., & Niemi, J. (2009). Exploiting degrees of inflectional ambiguity: Stem form and the time course of morphological processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 221.

Joanisse, M. F. & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: a connectionist model. *Proceedings of the National Academy of Sciences*, *96*, 7592–7597.

Juhasz, B. J. (2012). Sentence context modifies compound word recognition: Evidence from eye movements. *Journal of Cognitive Psychology*, *24*(7), 855–870.

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.

Just, M. A. & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 310.

Kaalep, H.-J. (1997). An Estonian morphological analyser and the impact of a corpus on its development. *Computers and the Humanities*, *31*(2), 115–133.

Kahnemann, D. & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, 1583–1585.

Karlsson, F. (1986). Frequency considerations in morphology. *STUF-Language Typology and Universals*, *39*(1-4), 19–28.

Karlsson, F. & Koskenniemi, K. (1985). A process model of morphology and lexicon. *Folia Linguistica*, *19*(1-2), 207–232.

Karttunen, L. (2003). Computing with realizational morphology. In *International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 203–214). Springer.

Kauschke, C. & Stenneken, P. (2008). Differences in noun and verb processing in lexical decision cannot be attributed to word form and morphological complexity alone. *Journal of Psycholinguistic Research*, *37*(6), 443–452.

Kazanina, N. (2011). Decomposition of prefixed words in russian. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1371.

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. R package version 0.3.0.

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.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, *44*(1), 287–304.

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*(1), 171.

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Proceedings of the 10th international conference on Computational Linguistics*, (pp. 178–181). Association for Computational Linguistics.

Kösling, K., Kunter, G., Baayen, R. H., & Plag, I. (2013). Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and speech*, 0023830913478914.

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

Kuchinke, L., Võ, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*(2), 132–140.

Kuperman, V. (2013). Accentuate the positive: Semantic access in english compounds. *Frontiers in psychology*, *4*.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*, 1089–1132.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*, 83–97.

Kuperman, V. & Dyke, J. A. V. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42 – 73.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2006). Morphological predictability and acoustic salience of interfixes in Dutch compounds. *JASA*, *122*, 2018–2024.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, *35*, 876–895.

Lõo, K., Järvikivi, J., & Baayen, R. H. (2017). Whole-word frequency and inflectional paradigm size facilitate Estonian case-inflected noun processing. *Manuscript submitted for publication*.

Lõo, K., Järvikivi, J., Tomaschek, F., Tucker, B. V., & Baayen, R. H. (2017). Production of Estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Manuscript submitted for publication*.

Laine, M., Niemi, J., Koivuselkä-Sallinen, P., & Hyönä, J. (1995). Morphological processing of polymorphemic nouns in a highly inflecting language. *Cognitive Neuropsychology*, *12*(5), 457–502.

Laine, M., Vainio, S., & Hyönä, J. (1999). Lexical access routes to nouns in a morphologically rich language. *Journal of Memory and Language*, *40*(1), 109 – 135.

Lavric, A., Clapp, A., & Rastle, K. (2007). Erp evidence of morphological analysis from orthography: A masked priming study. *Journal of Cognitive Neuroscience*, *19*(5), 866–877.

Lázaro, M., Illera, V., & Sainz, J. (2016). The suffix priming effect: Further evidence for an early morpho-orthographic segmentation process independent of its semantic content. *The Quarterly Journal of Experimental Psychology*, *69*(1), 197–208.

Leech, G., Rayson, P., & Wilson, A. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Lehtonen, M. & Laine, M. (2003). How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and Cognition*, *6*(03), 213–225.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38.

Lieber, R. (1981). Morphological conversion within a restrictive theory of the lexicon. In M. Moortgat, H. v. d. Hulst, & T. Hoekstra (Eds.), *The Scope of Lexical Rules* (pp. 161–200). Dordrecht: Foris.

Lieber, R. (1992). *Deconstructing Morphology: Word Formation in Syntactic Theory*. Chicago: University of Chicago Press.

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

Longtin, C., Segui, J., & Hallé, P. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *in press*, 0.

Longtin, C.-M. & Meunier, F. (2005). Morphological decomposition in early visual word processing. *Journal of Memory and Language*, *53*(1), 26–41.

Lõo, K., van Rij, J., Järvikivi, J., & Baayen, R. H. (2016). Individual differences in pupil dilation during naming task. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J. C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society, Austin, TX*, (pp. 550–555).

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.

Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, *28*(7), 905–916.

Marcus, G. F., Brinkman, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*, 189–256.

Marslen-Wilson, W. D., Bozic, M., & Randall, B. (2008). Early decomposition in visual word recognition: Dissociating morphology, form, and meaning. *Language and Cognitive Processes*, *23*(3), 394–421.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure*. London: Cambridge University Press.

McClelland, J. L. & Patterson, K. (2002a). Rules or connections in past-tense inflections: what does the evidence rule out. *Trends in the Cognitive Sciences*, *6*(11), 465–472.

McClelland, J. L. & Patterson, K. (2002b). 'words or rules' cannot exploit the regularity in exceptions: Reply to Pinker and Ullman. *Trends in the Cognitive Sciences*, *6*(11), 464–465.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one*, *12*(2), e0171935.

Milin, P., Filipović Durdević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 50–64.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. H. (2009). Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition* (pp. 214–252). Oxford: Oxford University 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 to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1271–1278.

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 Hebrew and Dutch. *Journal of Memory and Language*, *53*, 496–512.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*, 1–18.

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.

Niemi, J., Laine, M., & Tuominen, J. (1994). Cognitive morphology in Finnish: foundations of a new model. *Language and Cognitive Processes*, *9*, 423–446.

Papesh, M. H. & Goldinger, S. D. (2012). Pupil-blah-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics*, *74*(4), 754–765.

Pham, H. (2014). *Visual processing of Vietnamese compound words: A multivariate analysis of using corpus linguistic and psycholinguistic paradigms*. PhD thesis. University of Alberta, Canada.

Pham, H. & Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience*, *30*(9), 1077–1095.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. London: Weidenfeld and Nicolson.

Plag, I., Homann, J., & Kunter, G. (2015). Homophony and morphology: The acoustics of word-final s in english. *Journal of Linguistics*, 1–36.

Plaut, D. C. & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4/5), 445–485.

Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human, Perception and Performance*, *26*, 820–833.

Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, *58*, 1–21.

Ramscar, M. (2002). The role of meaning in inflection: Why the past tense doesn't require a rule. *Cognitive Psychology*, *45*, 45–94.

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

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the cost of learning, not cognitive decline. *Psychological Science*, *28*(8), 1171–1179.

Rastle, K. & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*(7-8), 942–971.

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.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 1090–1098.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255.

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

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., & Clifton Jr, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, *4*(3-4), SI21–SI49.

Robins, R. H. (1959). In defense of WP. *Transactions of the Philological Society*, *58*(1), 116–144.

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

Roelofs, A. (1996). Morpheme frequency in speech production: Testing WEAVER. In G. E. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 1996* (pp. 135–154). Dordrecht: Kluwer.

Rueckl, J. G. & Aicher, K. A. (2008). Are corner and brother morphologically complex? not in the long term. *Language and Cognitive Processes*, *23*(7-8), 972–1001.

Rumelhart, D. E. & McClelland, J. L. (1986a). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models* (pp. 216–271). Cambridge, Mass.: The MIT Press.

Rumelhart, D. E. & McClelland, J. L. (Eds.). (1986b). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. Cambridge, Mass.: MIT Press.

Schluroff, M., Zimmermann, T. E., Freeman, R., Hofmeister, K., Lorscheid, T., & Weber, A. (1986). Pupillary responses to syntactic ambiguity of sentences. *Brain and language*, *27*(2), 322–344.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, in press.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2017). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, in press.

Schmidtke, J. (2017). Pupillometry in linguistic research: an introduction and review for second language researchers. *Studies in Second Language Acquisition*, 1–21.

Schreuder, R. & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 131–154). Hillsdale, New Jersey: Lawrence Erlbaum.

Schreuder, R. & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Segalowitz, S. J. & Zheng, X. (2009). An erp study of category priming: evidence of early lexical semantic access. *Biological psychology*, *80*(1), 122–129.

Seidenberg, M. S. & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*(9), 353–361.

Selkirk, E. (1982). *The Syntax of Words*. Cambridge: The MIT Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

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

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

Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. (2011). Seeing a phrase time and again matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776.

Solomyak, O. & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042–2057.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition*, *18*(03), 419–437.

Soveri, A., Lehtonen, M., & Laine, M. (2007). Word frequency and morphological processing in finnish revisited. *The Mental Lexicon*, *2*(3), 359–385.

Spieler, D. H. & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, *15*(2), 225.

Sprenger, S. & van Rijn, H. (2013). Its time to do the math: Computation and retrieval in phrase production. *The Mental Lexicon*, *8*(1), 1–25.

Sproat, R. (1992). *Morphology and Computation*. Cambridge, Mass.: The MIT Press.

Stewart, T. & Stump, G. (2007). Paradigm function morphology and the morphology–syntax interface. In *The Oxford handbook of linguistic Interfaces*.

Stockall, L. & Marantz, A. (2006). A single route, full decomposition model of morphological complexity: Meg evidence. *The Mental Lexicon*, *1*(1), 85–123.

Stump, G. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press.

Stump, G. T. (1991). A paradigm-based theory of morphosemantic mismatches. *Language*, 675–725.

Stump, G. T. (1993a). On rules of referral. *Language*, 449–479.

Stump, G. T. (1993b). Position classes and morphological theory. In *Yearbook of Morphology 1992* (pp. 129–180). Springer.

Sun, C.-C. (2016). *Lexical Processing in Simplified Chinese: An Investigation Using a New Large-Scale Lexical Database*. PhD thesis, Eberhard Karls Universität Tübingen.

Tabak, W., Schreuder, R., & Baayen, R. H. (2005). Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence — Empirical, Theoretical, and Computational Perspectives* (pp. 529–555). Berlin: Mouton de Gruyter.

Tabak, W., Schreuder, R., & Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon*, *5*(1), 22–46.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes*, *9*(3), 271–294.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, *57A*, 745–765.

Taft, M. & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638–647.

Taft, M. & Forster, K. I. (1976a). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*, 607–620.

Taft, M. & Forster, K. I. (1976b). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*, 607–620.

Tomaschek, F. & Baayen, R. H. (2017). The consequences of lexical proficiency for articulation. *manuscript in preparation*.

Tomaschek, F., Tucker, B. V., Wieling, M., & Baayen, R. H. (2017). Vowel articulation affected by word frequency. In *Proceedings of 10th ISSP, Cologne*, (pp. 429–432).

Tomaschek, F., Wieling, M., Arnold, D., & Baayen, R. H. (2013). Word frequency, vowel length and vowel quality in speech production: an ema study of the importance of experience. In *INTERSPEECH*, (pp. 1302–1306).

Traficante, D. & Burani, C. (2003). Visual processing of Italian verbs and adjectives: the role of the inflectional family size. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 45–64). Berlin: Mouton de Gruyter.

Tremblay, A. & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

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

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

Ullman, M. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, *30*, 37–69.

van Rij, J., Baayen, R. H., Wieling, M., & van Rijn, H. (2016). itsadug: Interpreting time series, autocorrelated data using GAMMs. R package version 2.2.

Vannest, J., Bertram, R., Järvikivi, J., & Niemi, J. (2002). Counterintuitive cross-linguistic differences: More morphological computation in english than in finnish. *Journal of Psycholinguistic Research*, *31*(2), 83–106.

Vare, S. (2012). *Eesti keele sõnapered: tänapäeva eesti keele sõnavara struktuurianalüüs*. Eesti keele sõnapered: tänapäeva eesti keele sõnavara struktuurianalüüs. Eesti Keele Sihtasutus.

Viks, Ü. (1992). *Väike vormisõnastik: sissejuhatus & grammatika*. Number v1. Eesti Teaduste Akadeemia, Keele ja Kirjanduse Instituut.

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

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, *6*(9), e23613.

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences*, *109*(22), 8456–8460.

Wolter, B. & Gyllstad, H. (2013). Frequency of input and l2 collocational processing. *Studies in Second Language Acquisition*, *35*(03), 451–482.

Wood, S. N. (2006). *Generalized Additive Models*. New York: Chapman & Hall/CRC.

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

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

Yang, C. (2010). Three factors in language variation. *Lingua*, *120*(5), 1160–1177.

Yang, C. (2016). *The Price of Linguistic Productivity*. The MIT Press.

Young, F. A. & Biersdorf, W. R. (1954). Pupillary contraction and dilation in light and darkness. *Journal of comparative and physiological psychology*, *47*(3), 264.

Zekveld, A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: the influence of sentence intelligibility. *Ear and hearing*, *31*(4), 480–490.

Appendix A

Chapter 3: Supplementary materials

A.1 Appendix

Table A.1: Pairwise correlations between the key predictors in Experiment 1.

Predictor variables	1.	2.	3.	4.	5.	6.	7.
1. Whole-word frequency	1.00	0.42	0.20	0.08	0.30	-0.22	-0.36
2. Lemma frequency	0.42	1.00	**0.81**	0.23	**0.65**	-0.22	-0.23
3. Inflectional paradigm size	0.20	**0.81**	1.00	**0.71**	**0.62**	-0.09	-0.02
4. Inflectional entropy	0.08	0.23	**0.71**	1.00	0.34	0.05	0.01
5. Morphological family size	0.30	**0.65**	**0.62**	0.34	1.00	-0.25	-0.21
6. Orthographic neighbourhood density	-0.22	-0.22	-0.09	0.05	-0.25	1.00	**0.60**
7. Orthographic length	-0.36	-0.23	-0.02	0.01	-0.21	**0.60**	1.00

Table A.2: Pairwise correlations between the key predictors in Experiment 2.

Predictor variables	1.	2.	3.	4.	5.	6.	7.
1. Whole-word frequency	1.00	**0.58**	0.37	0.12	0.23	0.19	-0.25
2. Lemma frequency	**0.58**	1.00	**0.75**	0.31	0.38	0.17	-0.15
3. Inflectional paradigm size	0.37	**0.75**	1.00	**0.64**	0.30	0.13	-0.14
4. Inflectional entropy	0.12	0.31	**0.64**	1.00	0.11	0.02	-0.06
5. Morphological family size	0.23	0.38	0.30	0.11	1.00	0.18	-0.17
6. Orthographic neighbourhood density	0.19	0.17	0.13	0.02	0.18	1.00	**0.64**
7. Orthographic length	-0.25	-0.15	-0.14	-0.06	-0.17	**0.64**	1.00