Morphology (2018) 28:71–97
https://doi.org/10.1007/s11525-017-9318-7

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

Kaidi Lõo[1] · **Juhani Järvikivi**[2] ·
Fabian Tomaschek[3] · **Benjamin V. Tucker**[1] ·
R. Harald Baayen[3]

Received: 17 July 2017 / Accepted: 7 December 2017 / Published online: 15 January 2018

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. 2009a, 2009b). 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 a larger morphological fam-

This research was supported by the national scholarship program Kristjan Jaak, funded and managed by Archimedes Foundation in collaboration with the Estonian Ministry of Education and Research and by an Alexander von Humboldt research award (grant number 1141527) to the first and last author, respectively.

✉ K. Lõo
kloo@ualberta.ca

J. Järvikivi
jarvikivi@ualberta.ca

F. Tomaschek
fabian.tomaschek@uni-tuebingen.de

B.V. Tucker
bvtucker@ualberta.ca

R.H. Baayen
harald.baayen@uni-tuebingen.de

[1] Department of Linguistics, University of Alberta, 4-32 Assiniboia Hall, Edmonton,
AB T6G 2E7, Canada

[2] Department of Linguistics, University of Alberta, 4-55 Assiniboia Hall, Edmonton,
AB T6G 2E7, Canada

[3] Department of Linguistics, University of Tübingen, Wilhelmstrasse 19, 72074 Tübingen,
Germany

ily 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 (Niemi et al. 1994; Hankamer 1989; 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.

Keywords Whole-word frequency · Inflectional paradigm size · Estonian · Inflection morphology · Word naming

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 and 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 proportional 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 et al. 1999; Marcus et al. 1995; Pinker 1999; Taft and Forster 1976; 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 and 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 and Koskenniemi 1985; Koskenniemi 1984) and Estonian (Kaalep 1997), whereas in newer approaches, its role is much reduced (see e.g., Cotterell et al. 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 and Baayen 2008; Kuperman et al. 2009; Caselli et al. 2016), inflectional entropy (Milin et al. 2009b; 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 and Baayen 1997) co-determine lexical processing costs.

Importantly for the present study, Lõo et al. (2017) showed that lexical decision latencies for 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 basis of a large corpus. The present study was designed to replicate this finding in production, using a word naming task. We 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 first discuss the growing literature on whole-word frequency and paradigmatic effects in other, mostly morphologically less rich languages such as English and Dutch.

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 and Baayen 2008, 2012). Pham and 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 and 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, Niemi et al. (1994), Laine et al. (1995, 1999), Lehtonen and Laine (2003), 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 an implicit priming task, where participants first learned to associate word pairs and then produce the second word in the 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 effect.

This result was not replicated in a different paradigm by Janssen et al. (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, in another picture naming study with Dutch singular and plural nouns, Baayen et al. (2008) did not find whole-word frequency effects, which is consistent with Levelt et al. (1999). However, as reported below, they did find paradigmatic effects. 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 mega-study, more frequent compounds were read aloud with shorter production latencies. 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 and Baayen (2017) observed using electromagnetic articulography more co-articulation between the stem and inflectional affix in more frequent German verbs (see also Tomaschek et al. 2013, 2014). In summary, evidence is accumulating that also for speech production, whole-word frequency plays an important role.

This collection 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 and Marantz (2015) and 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 already 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 et al. 2012) and the Dutch Lexicon Project (Keuleers et al. 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.

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 (Moscoso del Prado Martín et al. 2004; Baayen et al. 2007; Tabak et al. 2005). In production, Baayen et al. (2008) and Tabak et al. (2010) found an opposite effect in Dutch picture naming: higher entropy correlated positively with response times (see also Bien et al. 2011).

Milin et al. (2009b), Baayen et al. (2011) 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 and 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 and Ernestus (2012). Kuperman et al. (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, thus 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 and 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, this result 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 all 14 cases of their singular and plural forms, but only the cases which make sense based on the meaning of the word. This point 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 the 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 a 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 actual use.

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

Table 1 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	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

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.

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 which may express subtle meaning differences,[2] the total number of forms for a single noun lemma can theoretically be well over 40 (see Table 1 *jalg* 'foot').

In addition to having a complex inflectional system, Estonian is characterized by productive derivation and compounding. Like Finnish (Moscoso del Prado Martín

[2] *jalgadel* and *jalul* both express adessive plurals, but *jalgadel* has more external locational meaning (something is *on the feet*) and *jalul* more a idiosyncratic meaning as 'back on the feet'.

Table 2 Part of morphological family for *jalg* 'foot'

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

et al. 2004), most Estonian words are part of large morphological families with up to a 1,000 members. Table 2 presents a fragment of the Estonian morphological family for *jalg* 'foot', which has over 300 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 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.

2 Experiment 1: Small-scale word naming study

2.1 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).

2.2 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.

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.

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.

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 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 et al. 2017 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):

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

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 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.[3] 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

[3] Hence, some inflected forms were counted neither as a part of inflectional paradigm nor orthographic neighbourhood count.

(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.

2.6 Analysis and results

Prior to the analysis, the data was cleaned and transformed (Baayen and Milin 2010). First, production latencies 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.

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 7 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 Generalized Additive Mixed Models (GAMM, Wood 2006; R-package *mgcv*, see also Baayen et al. 2017b). A standard Gaussian regression model predicts responses y_i as a function of a linear predictor η_i 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 and Tibshirani 1990; Lin and Zhang 1999; Wood 2006, 2011; Wood et al. 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 modelled 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 function, 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 modelled 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,

Table 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
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

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 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 modelled 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.

2.6.1 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 7 in Appendix A). 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 by manner of articulation. Further main or random effects did not reach significance. The complete model summary can be found in Table 3.

2.6.2 Acoustic durations

Acoustic durations decreased linearly for increasing whole-word frequency (see Fig. 1), and orthographic neighbourhood density. Neither experiment order nor stem

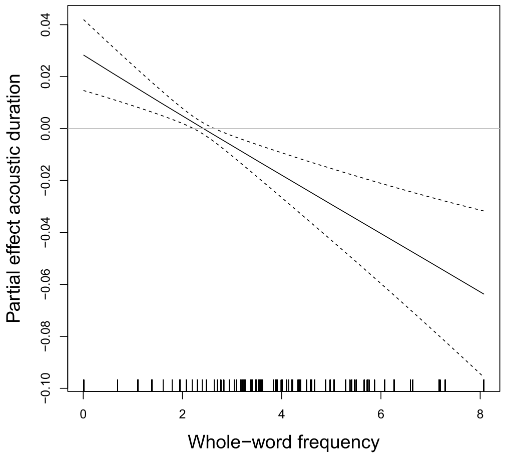

Fig. 1 Partial effects for whole-word frequency in acoustic duration of Experiment 1. The *solid horizontal line* represents the zero effect and *dashed lines* represent 95% confidence bands of the regression line for whole-word frequency

Table 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
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

complexity was again predictive of acoustic durations. They were tested 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 factors 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 4.

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 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 Experiment 2: Large-scale word naming study

3.1 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. 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 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.

3.2 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 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 participant's 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.

3.4 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 centre. 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 they were 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.

3.5 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. 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 (*powertransform-function* from the R-package *car* by Fox and 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.

3.5.1 Production latencies

A GAMM fitted to the log-transformed production latencies showed that increasing morphological family size (see the upper right panel of Fig. 2) linearly decreased production latencies. Response latencies increased with whole-word frequency. The upper left panel of Fig. 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 Fig. 2 shows this effect levels off as the paradigms increase in size. Finally, as orthographic length increased so did the production latency. The lower left panel of Fig. 2 shows that the effect of orthographic length 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 5.

3.5.2 Acoustic durations

Acoustic durations decreased linearly with increasing whole-word frequency (see the upper left panel of Fig. 3), and slightly non-linearly with inflectional paradigm size (see the upper middle panel of Fig. 3), orthographic neighbourhood density (see the lower left panel of Fig. 3) and non-linearly with morphological family size (see the upper right panel of Fig. 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,

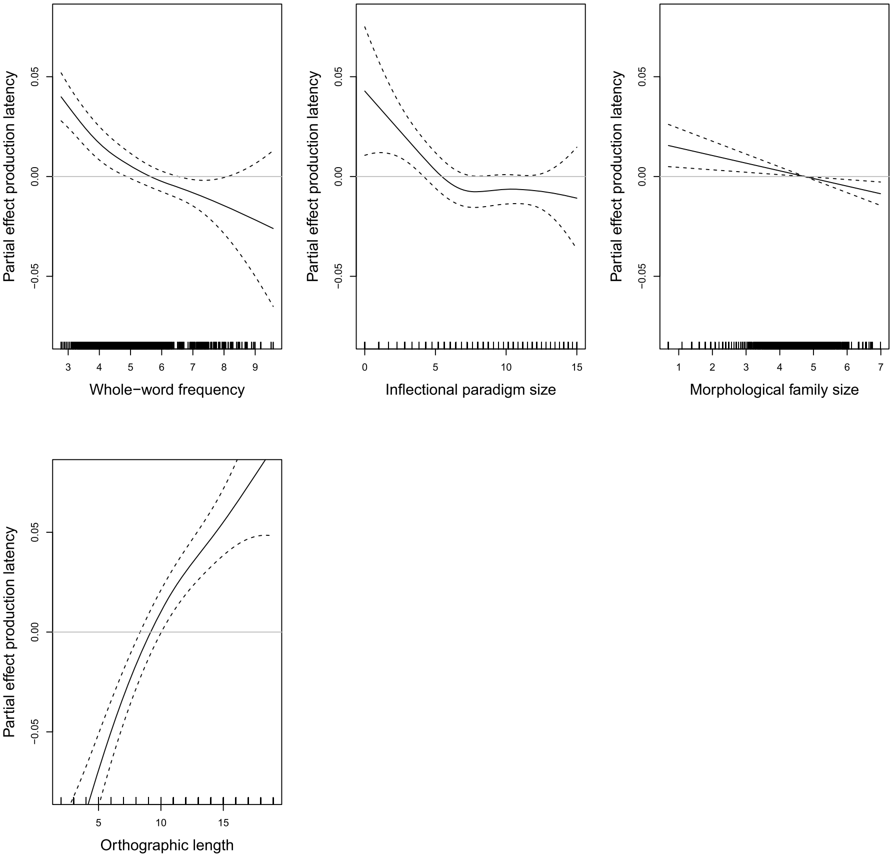

Fig. 2 Partial effects for whole-word frequency, inflectional paradigm size, morphological family size and orthographic length for the production latencies of Experiment 2. The *solid horizontal line* represents the zero effect and *dashed lines* represent 95% confidence bands of the regression line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

and manner of articulation. Further random effects did not reach significance. The complete model summary can be found in Table 6.

3.5.3 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

Table 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
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(Neighbourhood 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

response variable. We modelled the three predictors of theoretical interest as linear, while including trial and orthographic length as control variables as well as random intercepts for participant (adding item as a second random effect factor may lead to catastrophic data sparsity at extreme quantiles). Figure 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 for short production latencies at the .40 decile (see the left panel of Fig. 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 Fig. 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 frequency. 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.

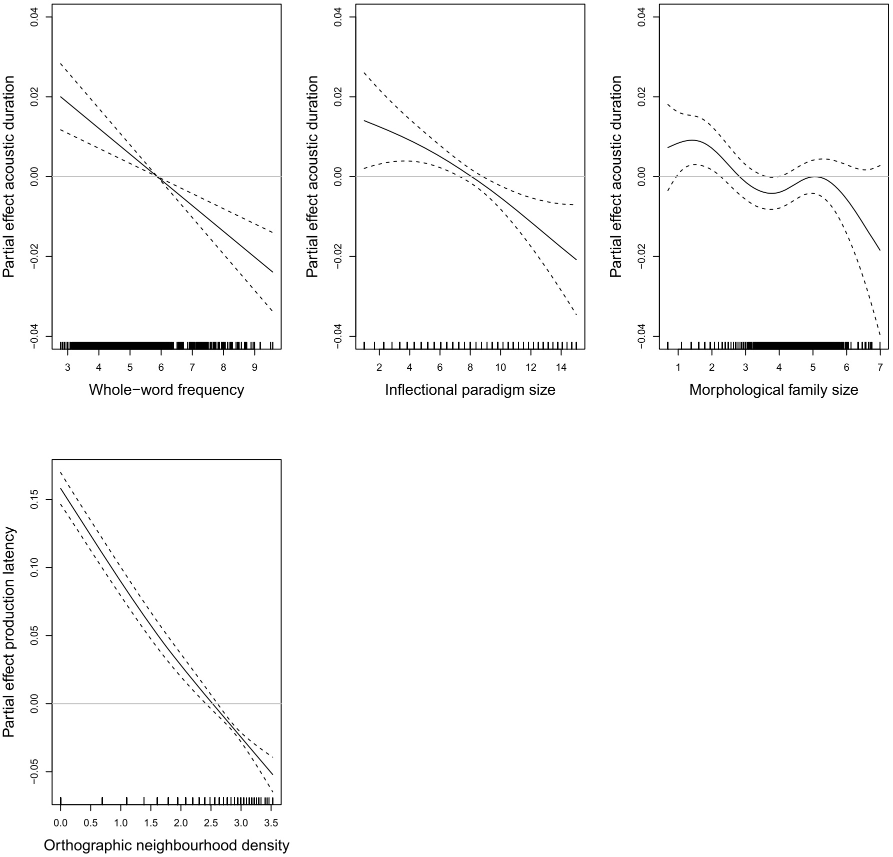

Fig. 3 Partial effects for whole-word frequency, inflectional paradigm size, morphological family size and orthographic density for acoustic duration of Experiment 2. The *solid horizontal line* represents the zero effect and *dashed lines* represent 95% confidence bands of the regression line for individual predictors

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.

Table 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(Neighbourhood density)	2.23	2.28	1003.61	<0.0001
s(Neighbourhood 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

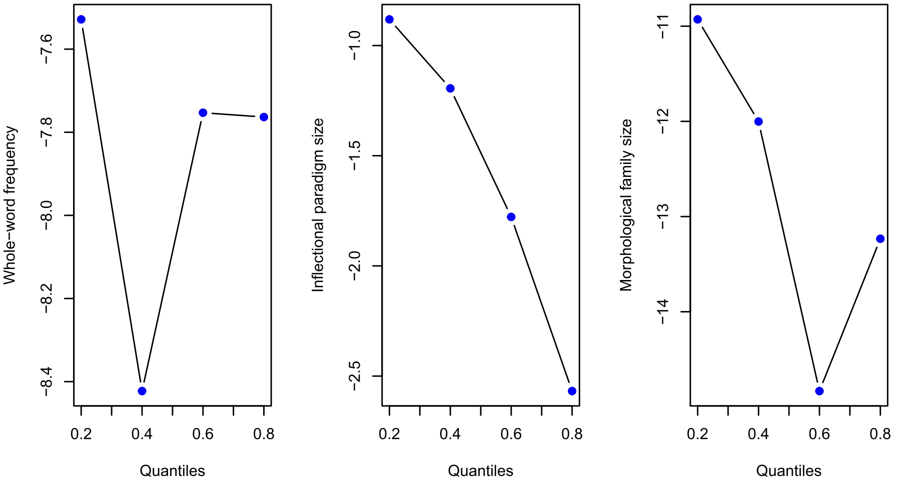

Fig. 4 Whole-word frequency, inflectional paradigm size, morphological family size β-coefficients in the second, fourth, sixth and eighth decile of production latency in Experiment 2

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, 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. (2009b) 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, contrary what happens in German (see Blevins et al. 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.

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 and 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 et al. 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 and 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 and 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 use (Shaoul et al. 2013, 2014).

Furthermore, Tremblay and 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 and 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 and van Rijn (2013) had participants produce Dutch clock time expressions, and found that more frequent expressions were produced faster. Arnon and 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.

4.3 General remarks

Although Distributed Morphology (Halle and 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 and 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, 2017a).

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 succinct 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 behavioural sciences, which continue to loop through the cycle of data collection, statistical analysis, and repeat.

> In the best kind of linguistic practice, simple 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.

Acknowledgements We would like to thank Pärtel Lippus and Einar Meister for their assistance with the data collection as well as the editor-in-chief Olivier Bonami and two anonymous reviewers for their helpful reviews on the earlier version of this paper.

Appendix A

Table 7 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. 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 8 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. 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

References

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

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), e0174,623. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174623.

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.

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

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., 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: de Gruyter.

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.

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

Baayen, R. H., Milin, P., Filipović-Đurđević, 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., Sering, T., Shaoul, C., & Milin, P. (2017a). Language comprehension as a multiple label classification problem. In *Proceedings of the 32nd international workshop on statistical modelling (IWSM)*. The Netherlands: Johann Bernoulli Institute, Rijksuniversiteit Groningen. 3–7 July, 2017.

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

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.

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.

Bien, H., Levelt, W., & Baayen, R. H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences of the USA*, *102*, 17,876–17,881.

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.

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

Blevins, J. P. (2013). Word-based morphology from Aristotle to modern WP. 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: 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*. Leiden: Brill. Chap. 8.

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

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. Preprint, arXiv:170609031.

De Jong, N. H. (2002). *Morphological families in the mental lexicon. MPI series in psycholinguistics*. Nijmegen: 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.

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

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd edn.). Thousand Oaks: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

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.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *Current studies in linguistics: Vol. 24. The view from building 20: essays in linguistics in honor of Sylvain Bromberger* (pp. 111–176). Cambridge: 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 W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 392–408). Cambridge: MIT Press.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. New York: 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 Tübingen.

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

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE*, *7*(3), 202, e3

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.

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.

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.

Keuleers, E. (2013). vwr: useful functions for visual word recognition research. http://CRAN.R-project.org/package=vwr, 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.

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

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2006). Morphological predictability and acoustic salience of interfixes in Dutch compounds. *Journal of the American Statistical Association*, *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. Human Perception and Performance*, *35*, 876–895.

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

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.

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.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, *61*(2), 381–400.

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.

Matthews, P. H. (1974). *Morphology: an introduction to the theory of word structure*. Cambridge: Cambridge University Press.

Milin, P., Filipović Durdević, D., & Moscoso del Prado Martín, F. (2009a). The simultaneous effects of inflectional paradigms and classes on lexical recognition: evidence from Serbian. *Journal of Memory and Language*, *60*, 50–64.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. H. (2009b). 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., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, *94*, 1–18.

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.

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.

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. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, *53*(1), 181–216.

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.

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.

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*.

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

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

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

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

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

Stump, G. (2001). *Inflectional morphology: a theory of paradigm structure*. Cambridge: Cambridge University Press.

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

Taft, M., & Forster, K. I. (1976). 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, in preparation). The consequences of lexical proficiency for articulation. Manuscript.

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).

Tomaschek, F., Tucker, B. V., Wieling, M., & Baayen, R. H. (2014). Vowel articulation affected by word frequency. In *Proceedings of 10th ISSP*, Cologne (pp. 429–432).

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: de Gruyter.

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.

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.

Vare, S. (2012). *Eesti keele sõnapered*. Tallinn: Eesti Keele Sihtasutus.

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*. Cambridge: MIT Press.