Cognition 175 (2018) 20–25

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Brief article

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

Kaidi Lõo[a,*], Juhani Järvikivi[a], R. Harald Baayen[b]

[a] University of Alberta, Canada
[b] University of Tübingen, Germany and University of Alberta, Canada

ARTICLE INFO

Keywords:
Whole-word frequency
Inflectional paradigm size
Estonian
Inflectional morphology
Semantic categorization
Generalized additive mixed-models

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.

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. Inflected forms (e.g., *tööd* 'works', *töös* 'at work', *tööga* 'with the work'), while inflected forms cluster into inflectional paradigms. Inflectional paradigms typically come with a few inflected variants, known as the principal parts, from which all other forms in the paradigm can be predicted (Blevins, 2006). Thus, 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, storing all word forms in a mental dictionary would exceed the storage capacity of the brain (Hankamer, 1989; Niemi, Laine, & Tuominen, 1994; Yang, 2016). Crucially, 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 been argued to play a fundamental role even in languages with simple morphologies, such as English (Fruchter & Marantz, 2015; Rastle, Davis, & New, 2004; Taft, 1994; Taft & Forster, 1975), it seems especially attractive for languages with rich morphology, such as Estonian, in order to minimize storage and maximize rule-driven computation (Pinker, 1999).

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;

* Corresponding author at: Department of Linguistics, 4-32 Assiniboia Hall, University of Alberta Edmonton, AB T6G 2E7, Canada.
E-mail address: kloo@ualberta.ca (K. Lõo).
[1] http://www.cl.ut.ee/korpused/grammatikakorpus/ (15.01.2017).

https://doi.org/10.1016/j.cognition.2018.02.002
Received 23 February 2017; Received in revised form 29 January 2018; Accepted 5 February 2018

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 effects 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 language 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 et al., 1994; Soveri, Lehtonen, & Laine, 2007; Vannest, Bertram, Järvikivi, & Niemi, 2002). One reason may be that inflection typically serves syntactic functions, such as grammatical role, number marking and agreement, whereas derivation and compounding tend to result in the formation of new words (see e.g., Booij, 2006, and for detailed discussion that the distinction between inflection and word formation is not an absolute one, Booij, 1993). 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 research 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 of the size of a word's morphological family (i.e., the count of derived words and compounds sharing a constituent) e.g., *worker*, *workforce*, *handwork*, while excluding inflectional variants from the counts for lexical processing have been studied extensively (Bertram, Baayen, & Schreuder, 2000; De Jong, Schreuder, & Baayen, 2000; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 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, Schreuder, & Baayen, 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, Milin, Filipović Đurđević, Hendrix, & Marelli, 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, Dijkstra, Schreuder, & Baayen, 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, Kostić, and Baayen (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, Filipović Durdević, & Moscoso del Prado Martín, 2009, see also Baayen et al., 2011 for prepositional entropy effects for English).

Table 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

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 1).

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 occur. Fig. 1 illustrates this point for Estonian. As a result of productive compounding, many Estonian nouns are attested with only one inflected case form. However, there are also nouns that have well-filled paradigms including variant forms, can comprise up to 38 different forms.

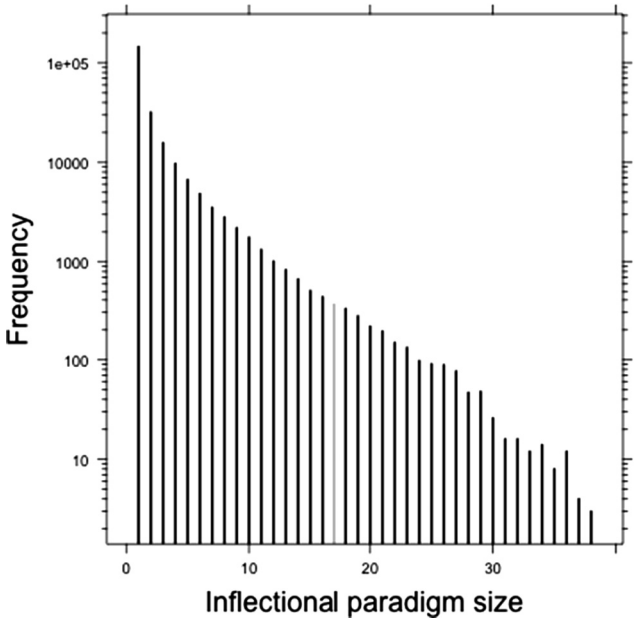

Fig. 1. Histogram of 231,891 noun paradigms in the Balanced Corpus of Estonian, 75% of which are paradigms of compound. The x-axis represents inflectional paradigm size, the y-axis shows the frequency of a particular paradigm size. Small inflectional paradigms are more common than large inflectional paradigms.