COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2017) 1–35

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12576

Using Statistical Models of Morphology in the Search for Optimal Units of Representation in the Human Mental Lexicon

Sami Virpioja,[a†] Minna Lehtonen,[b,c†] Annika Hultén,[d] Henna Kivikari,[d] Riitta Salmelin,[d] Krista Lagus[e,f]

[a]*Department of Signal Processing and Acoustics, Aalto University*
[b]*Department of Psychology, Abo Akademi University*
[c]*Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki*
[d]*Department of Neuroscience and Biomedical Engineering, Aalto University*
[e]*Faculty of Social Sciences, University of Helsinki*
[f]*Department of Computer Science, Aalto University*

Received 20 August 2016; received in revised form 25 October 2017; accepted 28 October 2017

Abstract

Determining optimal units of representing morphologically complex words in the mental lexicon is a central question in psycholinguistics. Here, we utilize advances in computational sciences to study human morphological processing using statistical models of morphology, particularly the unsupervised Morfessor model that works on the principle of optimization. The aim was to see what kind of model structure corresponds best to human word recognition costs for multimorphemic Finnish nouns: a model incorporating units resembling linguistically defined morphemes, a whole-word model, or a model that seeks for an optimal balance between these two extremes. Our results showed that human word recognition was predicted best by a combination of two models: a model that decomposes words at some morpheme boundaries while keeping others unsegmented and a whole-word model. The results support dual-route models that assume that both decomposed and full-form representations are utilized to optimally process complex words within the mental lexicon.

Keywords: Morphology; Mental lexicon; Statistical language modeling; Minimum Description Length principle; Unsupervised learning; Lexical decision; Word recognition; Psycholinguistics

Correspondence should be sent to Sami Virpioja, Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, P.O. Box 12200, FI-00076 Aalto, Finland. E-mail: sami.virpioja@aalto.fi

[†]These authors contributed equally to this work.

1. Introduction

A fundamental issue in cognitive science and psycholinguistics is the acquisition and representation of language and its grammar. Is language learning based on inherent constraints (e.g., Chomsky, 1965; Lidz & Gagliardi, 2015; Yang, 2004), and to what extent are general learning mechanisms capable of achieving the learning outcome (e.g., Ambridge & Lieven, 2015; Tomasello, 2003)? A related issue is whether the learned linguistic representations are grammatically structured or arise organically from statistical regularities of the input.

Morphology represents an area of language in which related words (e.g., clearly, unclear) bear systematic correspondences between form and meaning. Cognitive models of morphological processing have been developed to propose how these correspondences are encoded in our mental lexicons and whether and when during word processing we may utilize morphological information. The models have focused on whether morphologically complex words (inflected, derived, and compound words) are decomposed into their meaningful constituents, morphemes (e.g., clear+ly; Rastle & Davis, 2008; Taft & Forster, 1975), or processed as whole units (e.g., clearly; Butterworth, 1983; Hay & Baayen, 2005). These two single-route frameworks that assume only one type of representation have been challenged by dual-route alternatives (e.g., Diependaele, Sandra, & Grainger, 2005; Kuperman, Schreuder, Bertram, & Baayen, 2009; Niemi, Laine, & Tuominen, 1994; Schreuder & Baayen, 1995) which assume that both kinds of representations are possible.

A central theme in the research on morphological processing has been the balance between storage and computation. The question is whether it is more economical to store frequently co-occurring units as wholes or to compute them online, and where the limits for these two constraints are situated. The importance of chunking smaller elements and sequences into larger, integrated units is not only central in psycholinguistics but also more generally in cognitive science in topics such as memory and motor learning. Relevant for this discussion is the concept of optimization, that is, determining the most optimal units of representation, in terms of minimizing storage capacity and processing speed (see, e.g., Kuperman, Bertram, & Baayen, 2010; Schreuder & Baayen, 1995). Finnish, for example, is a morphologically rich language in which each noun has about 150 paradigmatic forms, and various clitic particles can additionally be attached to these forms. Storing all these word forms as whole units is thus unlikely to be economical for the storage capacity of the mental lexicon, suggesting that having them decomposed into morphological constituents is a useful strategy for the cognitive system. However, decomposition may entail a cost as well: Inflected Finnish words robustly elicit longer reaction times (RTs), larger error rates, and a greater number of eye-fixations than matched monomorphemic words (Hyönä, Laine, & Niemi, 1995; Laine, Vainio, & Hyönä, 1999; Soveri, Lehtonen, & Laine, 2007), suggesting that recognition of complex words is associated with a processing cost. By taking these two assumed limits into account, what is the most economical way to represent and process complex words? Here, we utilize computational models based on statistical learning and optimization to investigate the balance between storage (memorizing words as wholes) and computation (online decomposition and composition of word meanings) in the mental lexicon.

Computational models produce quantitative output that can be directly compared to continuous performance measures such as RTs in a word recognition task. If this kind of a model is able to successfully explain variation in a broad dataset measured using, for example, a word recognition task, it is likely that the way the model is built can tell us something essential about the cognitive processes in use when performing the task. Using computational models also forces one to be explicit about the kind of computations that give rise to these cognitive processes. Their quantitative nature makes them particularly well suited for investigating nuanced and graded (as opposed to categorical) effects that are likely to be relevant to the human cognitive system. In computational modeling, unsupervised statistical models utilize general learning principles to discover structure from the input and therefore mimic a situation in which the environmental input is central in learning of linguistic regularities such as morphology. Supervised models provide an interesting comparison point, as they, in turn, can be trained on pre-given linguistically structured input.

Computational models have been utilized to study a wide range of topics in language processing, such as word recognition building on the assumption that participants perform as Bayesian decision-makers (Norris, 2006), bilingual aphasia using self-organizing maps (Grasemann, Kiran, Sandberg, & Miikkulainen, 2011), and sentence processing with models based on either hierarchical or sequential sentence structure (Frank & Bod, 2011). With regard to morphological effects in word recognition, previous computational modeling research has not always taken morphemes as relevant units of processing. Instead, morphological effects have often been modeled, for example, by distributed-connectionist implementations (see, e.g., Rueckl, 2010; for a review) which assume that such effects can be explained by form-meaning regularities coded in the hidden units within the model. While connectionist models have succeeded in predicting some psycholinguistic effects and are grounded in the idea of neural networks capable of learning, their typical learning mechanism, back-propagation, has been criticized for its psychological and biological implausibility (O'Reilly, 1998, 2001).

Utilizing concepts from information theory in modeling lexical processing has, however, proved to be a promising approach, assuming processing costs of words to be proportional to the amount of information carried by them (see, e.g., Kostić, 1991; Milin, Kuperman, Kostic, & Baayen, 2009; Moscoso del Prado Martín, Kostić, & Baayen, 2004). The paradigmatic view has taken lexical words instead of morphemes as the basic linguistic units in the lexicon and assumes that lexical processing is influenced by probability distributions of inflectional paradigms and classes (Milin, Đurđević, & del Prado Martín, 2009; Milin, Kuperman et al., 2009; Moscoso del Prado Martín, Kostić, & Baayen, 2004) as well as morphological families (Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004; Schreuder & Baayen, 1997). Another, more recent amorphous approach is the Naive Discriminative Reader model (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Baayen, Shaoul, Willits, & Ramscar, 2016). It applies a principle of discriminative learning via a simple network structure that maps orthographic or phonetic input units directly to symbolic semantic units without hidden layers. The model does not incorporate morphemes or even words in its architecture but assumes that discriminative cues present in the visual input are enough to map the input to correct meanings (for a detailed description of this model, see Appendix S1). The present study, in contrast, focuses on computational models that start from the assumption that morpheme-like elements may be relevant units of representation within the mental lexicon.

One fundamental cognitive principle that is relevant within the computational language processing framework is the minimization of processing cost. According to the principle of least effort (Zipf, 1949), people expend the least effort possible in communicating a concept. For example, speakers often use economy in their articulation, which tends to result in phonetic reduction of speech forms. In computational linguistics, the principle of least effort has been captured, for instance, using the information-theoretic Minimum Description Length (MDL) principle (Rissanen, 1978, 1989). John Goldsmith, the inventor of Linguistica (Goldsmith, 2001), the first MDL-based learning model of morphology, describes the problem that a child faces when learning a language—including what are the words, their constituent morphemes, and the syntax of a language—as complex enough that only a suitable computational optimization approach could in principle solve it: “It seems to me that the only manageable kind of approach to dealing with such a complex task is to view it as an optimization problem, of which MDL is one particular style” (Goldsmith, 2001, p. 190). A computational optimization method might thus inform the selection of optimal units of representation of morphologically complex words in the mental lexicon. The MDL principle has previously been successfully applied to studying acquisition of grammar (Hsu & Chater, 2011).

We utilize a statistical model, Morfessor, that is inspired by the information-theoretic MDL principle. The model is trained in an unsupervised and language-independent manner, offering a description of how the learning of a morphology system might take place. This model was developed on the hypothesis that significant parts of morphological processing can take place via unsupervised learning.

We compare the model that attempts to find optimal lexical units in an unsupervised manner to a supervised model which is based on linguistically defined morphemes and to a model assuming whole-word representations. Using this approach, we aim to provide a view on the nature of the optimal units of representation within the human mental lexicon.

We compare the performance of statistical models by using psycholinguistic word recognition data that reflect the processing and storage cost of individual words in adults. Information theory provides a way to relate the probabilities given by statistical language models to the measures of cognitive processing cost of humans. Specifically, word recognition times are correlated to the self-information, or “surprisal,” which refers to the extent to which a word came unexpected to a reader or listener (Frank, 2013). Self-information has previously been studied, for example, in the context of sentence processing (Frank, Otten, Galli, & Vigliocco, 2015; Hale, 2001; Levy, 2008) and auditory word recognition (Balling & Baayen, 2012; Ettinger, Linzen, & Marantz, 2014). Self-information of a word is the negative logarithm of its probability estimated by a statistical language model. It can be considered as a cost of constructing or retrieving the word form: It corresponds to the minimum number of bits required to encode the word using the model. Typically self-information of a word is considered in its sentential context; here we consider it in the case of independent word forms. By utilizing self-information estimates, we can calculate cognitive prediction accuracy of the language model, that is, how well a language model is able to predict a measure of cognitive effort of word processing such as RT.

1.1. Morfessor

A specific computational model of morphology that we utilize is Morfessor (Creutz & Lagus, 2002, 2005a,b, 2007), in which learning is driven by the information-theoretic MDL principle. Morfessor has proven successful in various engineering tasks related to language, for example by improving speech recognition accuracy in strongly agglutinative languages such as Estonian, Finnish, and Turkish (Creutz et al., 2007; Hirsimäki et al., 2006). Although largely developed for engineering purposes, its initial inspiration came from cognition, viewing the brain as an efficient information-processing device that is likely to exhibit a principle of compact encoding of information. Morfessor creates a model of word structure based on observed words and analyzes the morphological structure of new words.

Morfessor learns agglutinative morphology without supervision, that is, without pregiven labels or feedback. It does not limit the number of morphs per word and is thus suitable for modeling complex morphology. While the probabilistic models applied by Morfessor can also be trained in a supervised manner with pre-segmented linguistic morphs as input, the main benefit of the method is its ability to learn segmentations of words from unannotated data. First, it stores word forms as wholes (assuming “one word is one morph”; e.g., *build*, *builder*). Then it can utilize these stored “morphs” in segmenting other incoming words: For example, after storing *build*, encountering *builder* will lead to storing also *-er* from *builder* separately, which can then be used in segmenting other words. The segmentation results are affected by the number of different morphs in the input (the different inflectional and derivational forms and compound words sharing the particular stem or affix). It searches for a segmentation which is simultaneously compact and provides an accurate description of the data.

Here, an accurate description can be considered as having a low average self-information (surprisal) over the words in the data. An extremely compact lexicon would include only letters, but it would provide a poor representation of the data, as the letters would need to be retrieved one by one. In contrast, an extremely accurate description would be provided by a lexicon of all the word forms in the data, but then the lexicon would be huge. Moreover, representing new word forms, or in other words, generalizability to new data, would then be a problem. The optimal balance between these two extremes is found by using a cost function based on the two-part coding scheme of the MDL principle by Rissanen (1978): The first part measures a cost of storage for the lexicon (a larger lexicon increases the cost), while the second part is related to the cost of computation of the data (more holistic units reduce the cost). Without the first part, all words would be stored only as whole units, hampering understanding of novel words consisting of the same morphemes.

1.2. The present study

Here, we apply Morfessor and the self-information estimates it produces to human morphological processing, and specifically address the controversial question of the optimal units of representation and processing in the mental lexicon. We aim to see whether the optimization principle that Morfessor utilizes leads to better correspondence with human word recognition RTs than other comparable, supervised models that build their lexicons on linguistically defined morphemes or solely on full forms. Morfessor is based on statistical morphs, and its default version allows some words to be segmented at their morpheme boundaries while keeping other morpheme boundaries unsegmented. We also specifically manipulate the emphasis the different Morfessor instances place on the cost of storage (full-form representations) versus the cost of computation (decomposition). This allows a closer evaluation within the same model type, on the optimal units of representation in the human mental lexicon.

Morfessor has previously been studied in a psycholinguistic setting by evaluating how well predictions of unsupervised Morfessor models correlated with the RTs for a set of monomorphemic and bimorphemic inflected Finnish nouns (Virpioja, Lehtonen, Hultén, Salmelin, & Lagus, 2011). The results were compared with predictions of letter-based *n*-gram models and a number of variables known to affect RTs. Our current study builds on this preliminary investigation, but considers a larger and more varied set of test words and uses mixed-effect regression modeling in the evaluation. We compare Morfessor to other statistical models which also produce self-information estimates, but have different underlying assumptions about the units of representation in the lexicon. For this purpose, we utilize supervised models (morph n-gram models) in which linguistically motivated morphological segmentations based on a morphological analyzer are given to the model. The performance of these morpheme-based models is compared to a word unigram model that only includes whole word forms.

Our hypotheses reflect the predictions that the two kinds of single-route models (full decomposition or full storage) versus dual-route models of morphological processing make about the mental lexicon. If, as the full decomposition models predict, the human word recognition system decomposes words exhaustively into morphemes and utilizes them as primary processing units (e.g., Taft, 2004), we should observe a high correspondence between values derived from morph *n*-gram models and RTs. Conversely, if the mental lexicon relies on full-form representations, RTs should correlate highly with self-information estimates of a word unigram model. Dual-route models, in turn, assume that the RT should be best predicted by the optimal units discovered by the unsupervised MDL principle (which may not always correspond to distinct linguistic morphemes), here implemented by the Morfessor Baseline model.

Crucially for the present question of optimal balance between storage and computation in morphological processing, Morfessor has an interesting but so far unexplored property that it enables manipulating the way in which the model emphasizes decomposition to morphemes versus full-form storage by settings of a single hyper-parameter in the model (Kohonen, Virpioja, & Lagus, 2010; Virpioja, Kohonen, & Lagus, 2011). A small value of the hyper-parameter provides a lexicon of short units (or "morphs" that the model stores), whereas a large value leads to a lexicon of long units. By varying the hyper-parameter, it is possible to investigate, within the same model type, the emphasis on full forms versus on decomposed parts that produces the best correspondence to human word recognition times.

Importantly, we also take into account the models' cross-entropy and complexity which both affect their performance in predicting word recognition RTs. Empirical cross-entropy, which is a standard evaluation measure for statistical language models in computational linguistics, estimates how unexpected a certain text corpus is with regard to the model trained by other text data (text prediction accuracy).

Cross-entropy is the average self-information (surprisal) over all words in the text, here over our stimulus words. Thus, it gives an estimate on the text prediction accuracy for the model. Humans have been shown to be effective in predicting linguistic material. For example, low cross-entropy values have been associated with a high accuracy in predicting reaction times in sentence processing (Fossum & Levy, 2012; Frank, 2009; Frank & Bod, 2011). Therefore, models that apply the principle of optimizing cross-entropy (for example, any statistical models that apply maximum likelihood or maximum a posteriori estimation) are likely to work better in predicting cognitive processing costs than models which do not have this feature built in them. We are not interested in models that improve cognitive prediction accuracy just by improving cross-entropy, as it is not likely to provide many new revelations regarding language processing of humans. Instead, if we have multiple models that are equally good at text prediction but use different internal representations and one is better at predicting reaction times of humans, it suggests that the representations included in that particular model are similar to those applied by humans.

To summarize, we use self-information estimates from computational models to investigate the optimal units of processing that adults use for recognition of morphologically complex words. We study whether the best correspondences to lexical decision RTs are provided by a supervised model based on linguistic morphs, by a model incorporating full form representations only, or by an unsupervised model that finds an optimal balance between these two alternatives. With these model comparisons, we aim to test dual- versus single-route models of morphology, that is, whether all words are exhaustively decomposed into morphological constituents or whether full-form representations are accessed for some or all complex words. We expect a comparison of these computational models to shed light on the optimal balance of storage and computation in human morphological processing.

2. Methods

In this section, we describe how the psycholinguistic experiment and the statistical models were set up and selected, and how the correspondence between the two was evaluated, using a regression model. We were interested in how well the different language models are able to predict the reaction times in psycholinguistic data sets (cognitive prediction accuracy). In the interpretations, we also took into account the models' cross-entropy, that is, text prediction accuracy, as well as the manner they segment the words compared to linguistic segmentations. The work flow and chronological order of the different steps are described in Fig. 1.

2.1. Psycholinguistic data

We applied two psycholinguistic data sets: the reaction time data reported by Lehtonen et al. (2007) and a new data set collected for the purpose of the present study. The former was used as a development set for selecting suitable parameters for the evaluated model types. The study was approved by the Institutional Review Board of the Center for

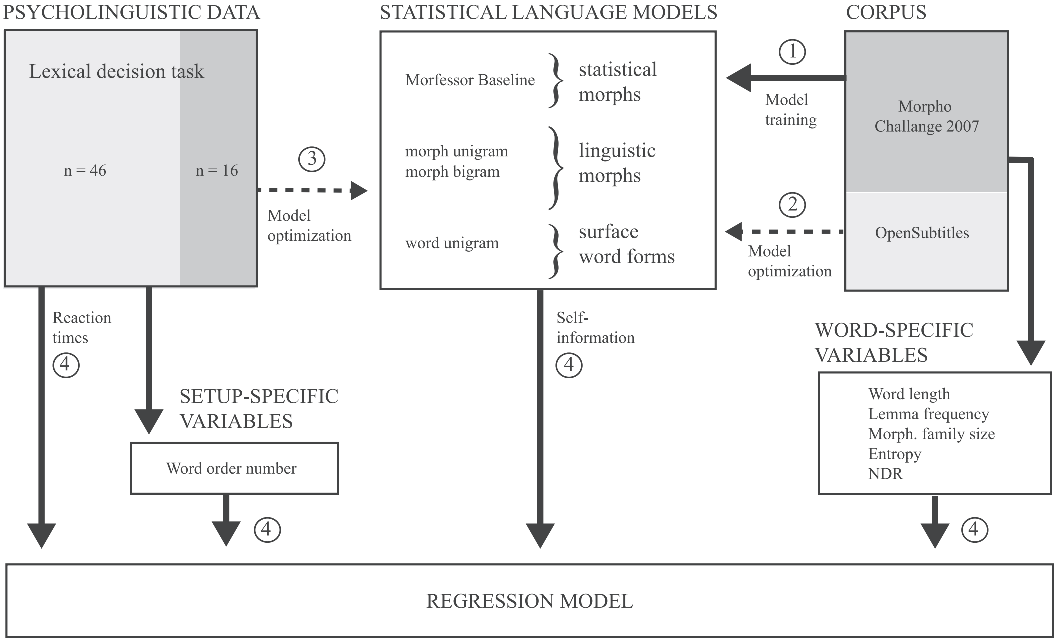

Fig. 1. Flowchart of the different stages of the analysis. The study contains two main sets of data: psycholinguistic data and data from statistical language models. All the models were trained (1) on the Morpho Challenge 2007 corpus, whereafter parameters for each model were optimized using novel corpus data (2) and a subset of the psycholinguistic data (3) not used in the final evaluation. The relationship between human morphological processing (quantified in terms of reaction times) and statistical language model performance (estimated in terms of self-information) was evaluated in a regression model together with a number of control variables that were both setup-specific and word-specific (4).

Cognitive Neuroscience, University of Turku, the Helsinki and Uusimaa Hospital District Ethics Committee, and the Aalto University Research Ethics Committee. All participants gave their written informed consent.

In the lexical decision study of Lehtonen et al. (2007), used as the development set, 16 native Finnish-speaking participants (8 females; mean age 24.7, SD: 2.39) were instructed to decide as quickly and accurately as possible whether a letter string was a real Finnish word or not, and to press the corresponding button. The stimuli included 320 real Finnish nouns composed of 80 high-frequency monomorphemic, 80 high-frequency inflected, 80 low-frequency monomorphemic, and 80 low-frequency inflected words, extracted from an unpublished Turun Sanomat newspaper corpus of 22.7 million word tokens by using a search program (Laine & Virtanen, 1996). The inflected words were bimorphemic. The words were interspersed by altogether 320 pseudowords. The elements in the pseudowords could include both real morphemes and pseudo-morphemes. The lengths and bigram frequencies (average frequency of letter bigrams) were similar for words and pseudowords. The letter-string length was 4–11 letters (mean 6.2, standard deviation 1.2). For additional details, please see Lehtonen et al. (2007).

For the new data set, 46 native Finnish speaking adults (38 females; mean age 27.2, SD: 6.5) participated in a lexical decision experiment similar to Lehtonen et al. (2007). They were recruited via university mailing lists. For 22 of these participants, eye-movements were also measured during the task; those data will be reported elsewhere. All participants reported having normal or corrected-to-normal vision and reported no language-related difficulties or neurological illnesses.

The stimuli in the new data set consisted of 300 unique nouns that were randomly selected from the list of word types in the Morpho Challenge 2007 corpus (Kurimo, Creutz, & Varjokallio, 2008). The same corpus was also used for training of the statistical models and analyzed by the morphological analyzer FINTWOL by Lingsoft, Inc. Words that were ambiguous or had a linguistically problematic analysis were replaced with new ones. Because of the Zipfian distribution of the words (Zipf, 1932), obtaining almost any high-frequency nouns in such a sample is unlikely. Thus, the sample was complemented with 60 randomly picked word forms of relatively high frequency. For the complete set, the word length was 4–16 letters (mean 10.3, SD: 2.8) and the number of morphemes 1–5 (mean 2.8, SD: 1.1). For other characteristics of the words, see Table 1. The morphemes included stems, inflectional and derivational suffixes, and clitic particles. Compound words were excluded from the sample. For the purpose of the task, we also included 360 pseudowords that followed the phonotactic rules of Finnish and had similar length to the real words.

Each trial began with an asterisk appearing in the middle of the screen for 500 ms, and the participants were to fixate their eyes on it. The asterisk was followed by a 500 ms blank screen, after which a stimulus item appeared. The item was visible for 1,500 ms after which an asterisk reappeared. The stimuli were divided into 6 blocks, with a short break between blocks. The order of the blocks was counterbalanced across participants, using a Latin Square. Before the experiment proper, a short practice session

Table 1
Statistics of control predictors and language model predictors over the stimulus words

Predictor	Range	Mean (SD)	Size	$\tilde{H}$	ρ	$-\Delta D$
Number of letters (log)	4–16	10.3 (2.8)	–	–	+0.625	182.8
Number of morphs (log)	1–5	2.8 (1.1)	–	–	+0.462	87.8
Surface frequency (log)	1–6,994	102.9 (548.7)	–	–	−0.595	156.8
Lemma frequency (log)	1–54,447	2,215.3 (5218.6)	–	–	−0.302	33.8
Morph. family size (log)	1–5,826	391.5 (791.4)	–	–	−0.255	23.2
Inflectional entropy	1.2–8.6	4.8 (1.7)	–	–	+0.279	28.5
Relative entropy	0.2–8.6	2.3 (2.1)	–	–	+0.573	41.2
NDR	0.8–18.4	8.3 (2.8)	–	–	+0.573	143.6
Word unigram	12.6–14.7	14.2 (0.6)	2.2×10^6	1.880	+0.596	157.5
Morfessor $\alpha = 10$	10.4–16.7	15.7 (1.1)	2.0×10^6	2.038	+0.616	170.8
Morfessor $\alpha = 5$	9.0–23.1	15.9 (1.9)	1.7×10^6	2.022	+0.542	125.3
Morfessor $\alpha = 2$	8.6–29.5	17.5 (3.9)	6.9×10^5	2.099	+0.526	118.0
Morfessor $\alpha = 1$	8.7–35.3	18.8 (4.5)	2.7×10^5	2.229	+0.647	195.0
Morfessor $\alpha = 0.8$	8.7–35.7	19.0 (4.5)	2.1×10^5	2.254	+0.659	205.8
Morfessor $\alpha = 0.5$	8.7–34.9	19.5 (4.7)	1.2×10^5	2.322	+0.640	191.0
Morfessor $\alpha = 0.2$	8.7–37.1	20.5 (5.1)	5.5×10^4	2.448	+0.619	175.7
Morfessor $\alpha = 0.1$	8.7–38.2	21.2 (5.5)	3.1×10^4	2.538	+0.597	159.5
Morfessor $\alpha = 0.05$	8.7–40.5	21.9 (5.8)	1.8×10^4	2.620	+0.588	154.1
Morfessor $\alpha = 0.02$	8.8–43.6	23.0 (6.4)	8.5×10^3	2.742	+0.577	146.8
Morfessor $\alpha = 0.01$	8.7–49.9	24.1 (6.6)	4.7×10^3	2.902	+0.557	135.3
Morph unigram	8.9–41.1	22.9 (5.9)	6.3×10^4	2.782	+0.567	141.7
Morph bigram	8.6–29.4	15.7 (2.8)	8.6×10^5	1.944	+0.620	176.3

Note. The columns show range, mean, and standard deviation of a variable, size and empirical cross-entropy $\tilde{H}$ of a language model, correlation ρ to the average reaction time, and decrease in deviance D for a regression model with random intercepts for participant and word, and word order number with subject-specific random slope as a control variable. All correlations are statistically significant ($p << .05$). Control variables with "log" have been transformed by the logarithmic function $\ln(1 + x)$ prior to estimating correlation and regression model.

(consisting of 16 stimuli not included in the actual experiment) was administered in order to familiarize the participants with the task.

As preprocessing in both data sets, we excluded all incorrect responses and reaction times of three SDs longer than each participant's mean; suspiciously short responses were not observed. The RTs to pseudowords were not included in the analyses. In the new dataset, data of two real-word items that had the same root ("sikaloitaan," "sikamaisuuttaan") and of two identical pseudoword items ("vihulkaisuuteen") were discarded from all analyses. For the remaining data, we took the logarithm of the reaction times.

2.2. Statistical language models

In addition to our primary model of interest, Morfessor, we evaluated three other statistical models: two so-called morph *n*-gram models (morph unigram and morph bigram) and a word unigram model. An overview of the properties of the different models is presented in Table 2. The details of the mathematical implementation of each model in the present study are reported in Appendix S1.

We focus on the simplest variant of the Morfessor methods, Morfessor Baseline (Creutz & Lagus, 2002, 2005b). It incorporates very little prior knowledge on human languages. We compare Morfessor to particular other statistical models that provide self-information estimates: a morph unigram model, a morph bigram model, and a word unigram model. The supervised morph unigram model is based on linguistically motivated morphological segmentations that are given to the model. This model has a similar structure as Morfessor as they both assume that morphs occur independently of one another; that is, a given morph is not predicted by the surrounding morphs. The morph bigram model is also on linguistic morphemes, but it predicts the upcoming morph on the basis of the previous one, and it has a more comparable cross-entropy with Morfessor than the morph unigram model. Finally, we compare these models to a word unigram model based on whole words, representing the costs associated with a lexicon of only full form representations.

The unsupervised Morfessor Baseline method (Creutz & Lagus, 2002, 2007) depends on the optimization of storage (as compact as possible) and an accurate description of the data. In accordance with the MDL principle, modeling is viewed as a problem of how to encode a data set efficiently in order to transmit it with a minimal number of bits. In contrast to the segmentation methods based on low-transitional probabilities between segment boundaries (Hafer & Weiss, 1974; Harris, 1955; Saffran, Newport, & Aslin, 1997), criticized by Baayen et al. (2016), Morfessor is not based on the local transitional probabilities but global probabilities of the segments. For example, although there is a low-transitional probability boundary between “pan” and “cake,” Morfessor trained on an English corpus is likely to keep the compound together as “pancake,” as p(pancake) is significantly higher than p(pan) × p(cake) and is thus supported by the MDL criterion. The same learning criterion can be used to find lexical constructions that consist of multiple words (Lagus, Kohonen, & Virpioja, 2009).

The outcome of the model optimization is dependent on the training data. In particular, increasing the size of the training corpus will produce a larger lexicon and longer lexical units (Creutz & Lagus, 2007; Virpioja, Kohonen et al., 2011). Let us assume that the initial corpus is doubled without entering any new word forms; that is, the same words are presented several times. This will double the cost of the computation (second part of cost

Table 2
Evaluated language models categorized by their units of representation and the structure of the statistical model

	Model Structure	
Model Units	Context-independent	Context-dependent
Statistical morphs	Morfessor Baseline	–
Linguistic morphs	Morph unigram	Morph bigram
Surface word forms	Word unigram	–

function) if the model parameters (i.e., the lexicon; first part of cost function) are not changed. The MDL criterion will balance the increase by favoring a more accurate model, which means including longer units in the lexicon. Longer lexical units mean an increase in the number of lexical units.

The size of the lexicon can be explicitly controlled by including a hyperparameter α, which modifies the weight of the training data in the optimized cost function and thus influences the size of the morphological units (Kohonen et al., 2010; Virpioja, Kohonen et al., 2011). That is, the parameter modifies the balance between the size of the lexicon and efficient description of data. By systematically manipulating the parameter α, we study how much the human cognitive system emphasizes a compact lexicon and processing efficiency in the human cognitive system. The extreme version of a compact lexicon would store all words as decomposed into single letters, whereas the most efficient processing would be achieved if all words are stored as holistic full form units. In Morfessor, a small value of α means that a greater number of observations are needed to store the input as it is encountered (e.g., as holistic units). Roughly speaking, the α parameter determines how sensitive the system is to storing repeatedly observed morpheme combinations.

With our training corpus, a large α value of 10 will result in long units corresponding to the full form word unigram model, whereas, a smaller value of 0.01 leads to decomposed units closest to linguistic morphs according to our investigations (see Fig. 2 and the section on linguistic segmentation accuracy below). Either these extremes or some value between the two may thus be able to capture the unit size that is most relevant to human word processing. We therefore trained several model instances between the extreme α values of 10 and 0.01 to study the effect on reaction time prediction.

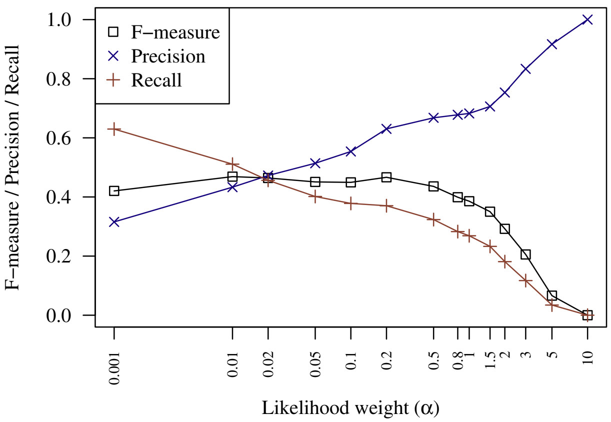

Fig. 2. Evaluation of Morfessor Baseline segmentations against linguistic reference segmentation. Increasing the likelihood weight α increases the length of the segments induced by the model, resulting in higher precision (fewer boundaries in within a linguistic morph) and lower recall (fewer boundaries between two linguistic morphs).

However, changing the weight hyper-parameter will also affect the size of the model. A large model inherently has a lower cross-entropy and therefore also a better text prediction accuracy (see Table 1). Since previous studies on sentence processing have implied that the text prediction accuracy may covary with the cognitive prediction accuracy (Fossum & Levy, 2012; Frank & Bod, 2011), we must make sure that any differences between the models' ability to predict human reaction times is not merely a direct consequence of the model size or cross-entropy. We therefore included cross-entropy as a relevant criterion in assessing theoretically interesting models with which to compare the Morfessor performance on predicting human reaction times.

The Morfessor model family is cognitively inspired and follows the principle of minimization of processing cost. We thus hope that the unsupervised learning algorithm of the Morfessor finds segmentations of words that have cognitive relevance. We know from previous studies on Morfessor (Creutz & Lagus, 2007; Virpioja, Kohonen et al., 2011) that the segmentation boundaries typically follow the boundaries of linguistic morphs, but that there are also differences. Thus, it makes sense to compare the performance of the Morfessor model to similar models trained in a supervised manner on linguistically motivated morphs.

The supervised morph unigram model determines the morphological unit based on rule-based linguistic morpheme borders. That is, the morph boundaries for the data set based on the morphological parser and the model parameters are estimated from the entire data set, before applying the model for the calculation of the probability for the words. An n-gram model is an $(n-1)$th order Markov model. The unigram model ($n = 1$) therefore assumes that the units occur independently and is in this respect similar to the Morfessor Baseline model. However, the unigram model has a larger cross-entropy than Morfessor at any α value above 0.02 (Table 1). A better comparison to Morfessor in this respect is offered by the morph bigram model which has more comparable cross-entropy values but is still based on linguistically defined morphemes. In this model, morphs are not context-independent: A given morph is predicted by the previous one.

The final model, the word unigram model, was chosen as it captures the probability of a word form as estimated based on its frequency in the training data. This model entails no morphological segmentation, as each word is represented in its full form.

2.2.1. Model training and optimization

The language models were trained on the Morpho Challenge 2007 data set distributed in the Morpho Challenge 2007 competition (Kurimo et al., 2008), available from http://morpho.aalto.fi/events/morphochallenge/. The Finnish-language corpus has been collected from World Wide Web as a part of the Wortschatz collection (Quasthoff, Richter, & Biemann, 2006). The number of word types in the corpus is 2,206,719, and the number of word tokens is 44,076,925.

The morphological analysis required by the supervised morph n-gram models was performed by the morphological analyzer FINTWOL by Lingsoft, Inc. It applies the two-level morphology model by Koskenniemi (1983). FINTWOL was able to analyze 1.7 million of the 2.2 million words. The rest of the words—mostly proper nouns, foreign words, and misspelled words—were discarded. The analyzer is highly accurate; for example, the 360 word forms in our development set, selected independently of the analyzer's output, were all correctly analyzed. The analyses were further processed using the automatic tools by Creutz and Lindén (2004) to obtain both linguistic morphemes and the corresponding segments (morphs) for each word.

The smoothing methods for the *n*-gram models, described in Appendix S1 of the Supplementary Material, were selected based on the psycholinguistic development set. For the optimization of the discount parameters used by the smoothing methods we used a corpus consisting of the Finnish subtitles of a single movie (*High Fidelity*; 2,478 word types and 6,614 word tokens) from the OpenSubtitles corpus collected from http://www.opensubtitles.org/ (Tiedemann, 2009). The development corpus has to be small for computational efficiency, and selecting a domain that differs from the training corpus helps to avoid overfitting.

It is not evident whether the language models for individual word forms should be trained on word tokens, on word types, or on something in-between. The Morfessor Baseline model extracts segments that correspond more closely to linguistic morphs when trained on types than tokens, as many inflected high-frequency words will not be segmented in the latter case (Creutz & Lagus, 2005b). An example of interpolating between types and tokens is application of a Pitman-Yor process to adapt the distribution of word types into the observed token count distribution (Goldwater, Griffiths, & Johnson, 2006, 2011).

We took a more straightforward approach and applied a logarithmic function $f(x) = \ln(1 + x)$ to dampen the effect of the counts. This way of dampening improved the reaction time prediction on the psycholinguistic development set when assessed by the word unigram model. We did not select the dampening separately for each language model, because then we would not have been able to fairly compare the cross-entropies of the models.

2.2.2. *Linguistic segmentation accuracy for the Morfessor model*

In order to investigate in more detail what kinds of units the unsupervised Morfessor Baseline produced in our set of items, we assessed its segmentation performance in light of linguistically correct segmentations. Accuracy of a morphological segmentation is typically estimated by calculating the precision and recall of the segmentation boundaries (Hafer & Weiss, 1974). Precision and recall are usually combined by taking their harmonic mean, which is called an F-measure. The F-measure was used to find the instance of Morfessor in which the segmentations were closest to a linguistic analysis.

As an example, consider possible segmentations for word *segmentations*. There are 12 possible boundaries between the letters. Two boundaries can be considered linguistically correct (*segment+ation+s*). Given a predicted (non-linguistic) segmentation, precision is the ratio of correct boundary predictions ("true positive") to all the predicted boundaries ("positive"), while recall is the ratio of correct predictions ("true positive") to all correct boundaries ("true"). If our prediction was *seg+men+ta+tion+s*, precision would be 1/4, recall 1/2, and F-measure $(2 \times 1/2 \times 1/4)/(1/2 + 1/4) = 1/3$.

We calculated linguistic segmentation accuracy of the Morfessor Baseline model for the stimulus words. Fig. 2 shows the precision, recall, and F-measure of model as function of the likelihood weight parameter α. According to the F-measure, the segmentation is closest to the linguistic segmentation at $\alpha = 0.01$, but the steepest decrease in F-measure and recall starts when α increases above 1.

Our linguistic analysis of the corpus based on FINTWOL also indicates the functional types of the morphs, and we use those to provide further automatic analysis of the segmentations indicated by Morfessor. The morphs in our stimuli include stems (STEM), derivational suffixes (DERIV), inflectional suffixes that consist of case inflections or possessive suffixes (INFL), and clitics (CLITIC). Each proposed segmentation boundary that occurs inside of a morph lowers precision; we will call this a disparity in precision. Each missed segmentation boundary between two morphs lowers recall; we will call this a disparity in recall. Thus we can calculate precision disparities for each of the four morph types and recall disparities for each ordered pair of morph types (STEM+DERIV, STEM +INFL, DERIV+INFL, etc.). Given our example (*segment+ation+s*), it contains a stem, a derivational suffix, and an inflectional suffix. Thus, the prediction *seg+men+ta+tion+s* would receive two precision disparities for STEM, one precision disparity for DERIV, and one recall disparity for STEM+DERIV.

2.2.3. Self-information as a predictor of human reaction times

In order to determine how an arbitrary probabilistic model such as *n*-gram model or Morfessor should relate to the human reaction times, the processing cost for each word in the respective models needs to be quantified. The self-information or surprisal of a word, $-\log p(w)$, has been shown to correlate strongly with the cognitive load when a word is processed in context (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Frank, 2009; Hale, 2001; Levy, 2008; Wu, Bachrach, Cardenas, & Schuler, 2010). Self-information is also directly related to the common frequency statistics used in psycholinguistic experiments: For example, the logarithm of the surface frequency of word w in a corpus is simply an unnormalized self-information from a unigram language model estimated from the same corpus. Accordingly, we made the assumption that the reaction time for a word is linearly proportional to the self-information of the word estimated by a probabilistic model (e.g., Smith & Levy, 2013).

For language models based on morph-like units, there may be several ways to split one word form into the units. Then the probability $p(w)$ is actually the sum of probabilities over all possible segmentations of w:

$$p(w) = \sum_{m_1 \ldots m_n = w} p(m_1 \ldots m_n) \tag{1}$$

However, this would include ungrammatical segmentations even for the supervised models based on linguistic morphs. For example, the English word *stairs* could be segmented ungrammatically to linguistic morphs *st* + *air* + *s*, where *st* is a common superlative suffix for adjectives. Instead of marginalizing over the segmentations in Eq. (1), we used the probability of a single segmentation. For the morph *n*-gram models, we took the grammatically correct segmentation. For the unsupervised Morfessor models, we found the most likely segmentation with an extension of the Viterbi algorithm and used the probability of that segmentation. We also tested the sum over all possible segmentations, but that yielded worse reaction time predictions on the development set.

2.3. Comparing statistical language models and psycholinguistic data

2.3.1. Regression models

To evaluate how well the language models can predict the reaction times of the test participants, we used mixed-effect multiple regression (for an introduction, see Baayen, Davidson, & Bates, 2008). We applied the lme4 R software package (Bates, Mächler, Bolker, & Walker, 2015); for details, see Appendix S2 of the Supplementary Material. In general, we studied whether the prediction of a certain language model could improve the regression model in the presence of centered control predictors. The improvement over the baseline model (i.e., control predictors only) was measured with the decrease in deviance. The likelihood ratio test was applied to test whether the nested model improved significantly over the baseline model. For the comparison of two different language models, we considered the Akaike information criterion (AIC) of the regression models; a smaller value means a better quality of the regression model.

We studied several regression models with an increasing number of control predictors. As we are interested in how the language models alone can predict human word recognition, we first did not include any word-specific control predictors. However, we did include a setup-specific control predictor accounting for the order in which the stimulus words were presented to the participants.

In the second test we added the word unigram predictions as a control variable. This was done in order to evaluate if language models based on linguistic or statistical morphs would improve the predictions only because they approximate the self-information based on the surface frequencies.

In order to investigate whether the language models contribute anything additional to the known psycholinguistic variables, we next considered regression models that incorporated a whole range of word-specific variables known to affect human word recognition in addition to the word unigram model (see list below). We looked for the combination of these variables and their two-way interactions that would provide the best baseline regression model, measured by the lowest AIC. We selected the control variables with a greedy search: Instead of testing all possible combinations, which would be computationally difficult, we added one predictor (or interaction of two predictors) at a time, retained the one that yielded the largest improvement for the current regression model, and continued until there was no further improvement.

Finally, we considered which of the original word-specific control predictors would improve the regression model result if the language model prediction was already in use. The aim was to study the relationships between each language model and these known variables, for example, to see how much of the effects of these variables are incorporated in our language models. We took the regression model of the first test as a baseline model, and for each of the word-specific variables, we used the likelihood ratio test to see whether it provided any further improvement.

2.3.2. Control predictors

Word order number: As the average reaction times tend to change according to how many stimulus words the participant has already seen, we added the logarithmically transformed presentation order number of the word for each participant as a setup-specific control predictor.

Word length: Two measures of word length were included: the number of letters and the number of morphemes.

Lemma frequency: Lemma frequency is the summative frequency of all the inflectional variants of a single stem (e.g., Baayen, Dijkstra, & Schreuder, 1997; Bertram, Baayen, & Schreuder, 2000; Taft, 1979) and assumed to affect the speed of accessing the stem when decomposing a complex words.

Morphological family size: Morphological family size is the number of derivations and compounds where the noun occurs as a constituent (e.g., Bertram et al., 2000; Moscoso del Prado Martín, Bertram, et al., 2004; Schreuder & Baayen, 1997). As such, it is considered a measure of lexical interconnectivity between morphologically related words. Complex words with a large family size have been shown to be processed faster than those with small morphological families (Bertram et al., 2000).

Paradigmatic entropy: Paradigmatic entropy (Kostić, 1991; Milin, Kuperman et al., 2009; Moscoso del Prado Martín et al., 2004) is operationalized as two different variables, which are based on the assumption that processing of a word is influenced by the amount of information in its inflectional paradigm and inflectional class (Milin, Đurđević, & del Prado Martín, 2009). Inflectional entropy is the expected amount of information load in an inflectional paradigm. The lexical units with a higher information load are assumedly more costly to retrieve. The more balanced the frequency distribution of the inflected variants within a paradigm is for a word, the higher the entropy. However, this variable has been reported to show facilitatory effects in lexical decision (Baayen, Feldman, & Schreuder, 2006). Relative entropy, in turn, measures the divergence between the distribution of the word's inflectional paradigm and the frequency distribution of the case endings for the inflectional class of the word (the set of words that are inflected in the same way). Lexical processing costs have been shown to be larger the greater the divergence between these distributions (Milin, Đurđević, & del Prado Martín, 2009). For implementation details, see Appendix S1 in the Supplementary Material.

Naive discriminative reader (NDR): The naive discriminative reader is a two-layer network based on the Rescorla-Wagner model (Rescorla, 2007; Rescorla & Wagner, 1972) that associates a set of cues with a set of outcomes. It has been proposed as an amorphous model of morphological processing (Baayen et al., 2011, 2016). Following Baayen et al. (2011), we used letter unigrams and bigrams as cues, and morpheme labels from a morphological analysis as outcomes (see Appendix S1 for details). Given the input cues, activation of the correct outcome, relative to the activations of competing outcomes, is used for predicting the reaction times.

2.3.3. Factors affecting the interpretation of the regression models

When comparing different statistical language models to each other, one needs to account for the inherent properties of the models such as the model's text prediction accuracy and model size. The size or complexity of the models provides an idea of how accurate predictions of the text data can be expected. We defined the model size as the number of non-zero probability estimates stored by the model. For unigram models, including Morfessor Baseline, it is the size of the lexicon. For *n*-gram models, it is the total number of *n*-grams. The model sizes are reported in Table 1.

The accuracy of the statistical language models with respect to the text data can be measured with cross-entropy. We used the empirical cross-entropies $\tilde{H}$ calculated over the psycholinguistic test set (for details, see Appendix S1). The words were weighted with log-dampened frequencies in the same manner as in training the language models. The use of the same dampening as for training the models is important, as otherwise the cross-entropy measure would not correspond to the maximum likelihood optimization criterion used in training. As empirical cross-entropy is a weighted average of the self-information estimates $-\log p(w)$ over the test set words, the smaller the self-information estimates, the better the text prediction accuracy of the model. However, the self-information values are bounded by the fact that that $\sum_{w \in T} p(w) \leq 1$ for any set of words T. As we predict also the reaction times with the self-information values, accurate estimates of self-information for the test set words, indicated by a low $\tilde{H}$, should generally improve the outcome of the regression models. The cross-entropy values are shown in Table 1.

3. Results

3.1. Deviance in the regression model versus cross-entropy

In our evaluations, we focused on four models: Morfessor Baseline, morph uni- and bigram models, and a word unigram model. In our first regression test, the baseline regression model included only word presentation order number as a control predictor. The self-information estimates of the language models were added to this model and the decrease in deviance; that is, the improvement in the prediction ability compared to the baseline regression model, was measured. The decreases in deviance with respect to the cross-entropies of the language models are shown in Fig. 3. The baseline regression model and the coefficients and *p*-values for the language model predictors are presented in Appendix S2 (Tables B1 and B2).

Our results show that the best Morfessor model instance outperformed both of the morph n-gram models, based on linguistically defined morphs, despite a more favorable cross-entropy of the morph bigram model. Moreover, this Morfessor model instance also performed better than the word unigram model that had the best cross-entropy value of all the models studied. When comparing the different Morfessor model instances with one another, the best prediction accuracy was found at the likelihood weight value $\alpha = 0.8$: It performed better than the model instance at $\alpha = 10$, based on whole words, as well as the instance providing units closest to linguistic morphs ($\alpha = 0.01$; see Figure 2). The difference between the word unigram and Morfessor at $\alpha = 10$ results are due to the smoothing method applied in the unigram model (see Appendix S1).

With regard to the effect of cross-entropy on cognitive prediction accuracy, the general pattern of results follows the tendency that has been observed previously in sentence processing (Fossum & Levy, 2012; Frank & Bod, 2011): Text prediction accuracy covaries with cognitive prediction accuracy. However, a few interesting exceptions are observed.

When increasing the value of α, Morfessor had good cognitive prediction accuracy with respect to its cross-entropy until the conspicuous drop in the accuracy for the model with a likelihood weight at 2. An explanation for the drop is revealed by further analysis of the self-information values (Fig. 4). With this value, an approximately even number of words are represented as a whole (single fragment) or by two fragments. Given that the self-information is based on the product of the fragment probabilities, doubling the number of fragments means a large increase in self-information. While both sets are individually well correlated with the reaction times, the difference between their average self-information is so large that it lowers the correlation over the whole set of words. This sets a limitation to exact interpretations of the cognitive prediction accuracy of the model with α values at this range.

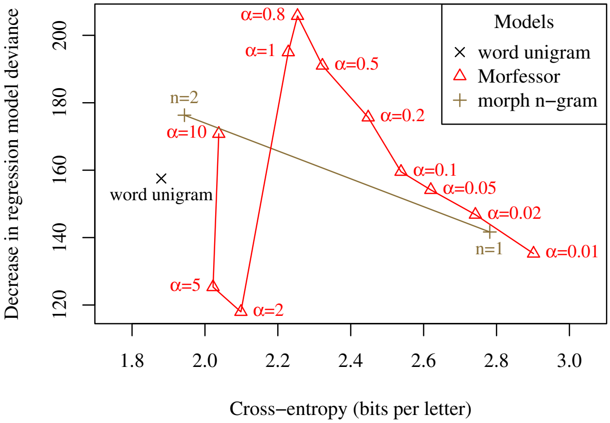

Fig. 3. Cognitive prediction accuracy (decrease in deviance) versus text prediction accuracy (cross-entropy) of the language models. The improvement over the baseline regression model is statistically significant ($p_{\text{anova}} \leq .05$) for all models.

Beyond the likelihood weight value of 2, the performance of Morfessor did not rise close to the level of Morfessor at α = 0.8 even at the lowest cross-entropy values at α = 10, that is, the level at which the predictions were based on full word forms. While both Morfessor at α = 10 and the word unigram model performed moderately well in predicting RTs in the present study, they also had the lowest cross-entropy values and still did not outperform the Morfessor instances with lower values of α.

In the first stage of analyses, the regression model included only one estimate for self-information at a time. The next step was to check if any of the language models could improve the regression model even when the typical estimator of self-information, surface frequency, was included as a control variable. We added the predictions from the word unigram model as both a fixed effect and a participant-specific random effect. The results are shown in Fig. 5; the details of the regression models are given in Appendix S2 of the Supplementary Material (Tables B3 and B4). All language models still yielded significant improvements. However, the Morfessor Baseline models with high values of α as well as the morph bigram model performed relatively worse than the other language models despite their lower cross-entropies.

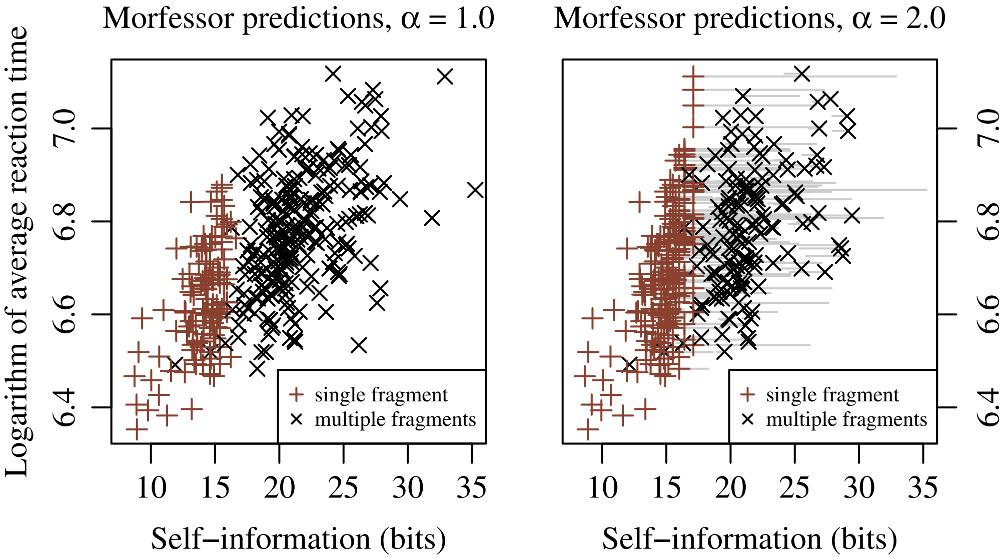

Fig. 4. The effect of using a large likelihood weight for Morfessor. Each point represents a stimulus word. When the likelihood weight α is increased from 1.0 (left) to 2.0 (right), the words are more clearly divided into two clusters by their self-information. The change of the self-information of individual words between 1.0 and 2.0 is shown by the light gray lines in the right-side scatter plot. With a high value of α (e.g., 10), in practice all words are stored in the lexicon as whole. When the value of α is low, only the most frequent words are unsegmented, and many words are segmented to more than two lexical items. In between (here at α = 2.0), there is the case where a considerable part of the words are encoded directly in the lexicon, while other, slightly less frequent words, are still segmented, mostly into two fragments. The difference in average self-information between the words that are represented as whole forms and split into two or more fragments is so large that it dominates the variance of self-information over the stimulus words. This lowers the correlation over the whole set of words even if the two correlation was high for both subsets separately.

Next we studied how the models used here relate to commonly investigated psycholinguistic variables and selected an optimal set of the additional predictors and their two-way interactions by a greedy search as explained in Section 2.3.1. The initial regression model included only the word order number and word unigram predictions. The coefficients of the baseline regression model after the search are shown in Table B5 of Appendix S2. Then we again computed nested regression models with each remaining language model. The decreases in deviance are shown in Fig. 6 and details of the regression models in Table B6 of Appendix S2. This time only the morph bigram model and Morfessor at $\alpha = 0.8$, 1, 5, and 10 provided significant improvements.

3.2. Language models as control predictors

For the final regression model evaluation, we took combinations of a language model predictor and each of the word-specific variables. Now the self-information values from a language model were added as a control predictor to the first baseline regression model (Table B1 in Appendix S2), including random slopes for the predictor. Then each of the other variables were tested for any further improvement in regression model accuracy. The results for word unigram, the best-performing Morfessor Baseline instance ($\alpha = 0.8$), and morph bigram model are shown in Table 3. The results show, for example, that the Morfessor model instance explains the same variance as variables related to morphology, such as lemma frequency and morphological family size, but not the same variance as NDR.

3.3. Linguistic segmentation accuracy

The segmentations of the Morfessor models were compared against linguistic morph segmentations to get quantitative assessment for the segmentation. As explained in the Methods

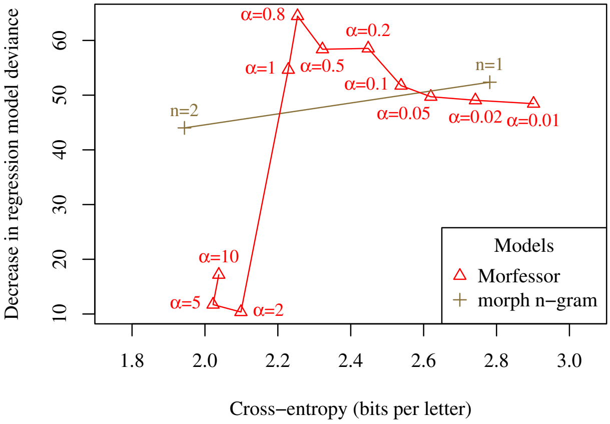

Fig. 5. Cognitive prediction accuracy (decrease in deviance) versus text prediction accuracy (cross-entropy) of the language models; word unigram as a control predictor. The improvement in the baseline regression model is statistically significant ($p_{\text{anova}} \leq .05$) for all models.

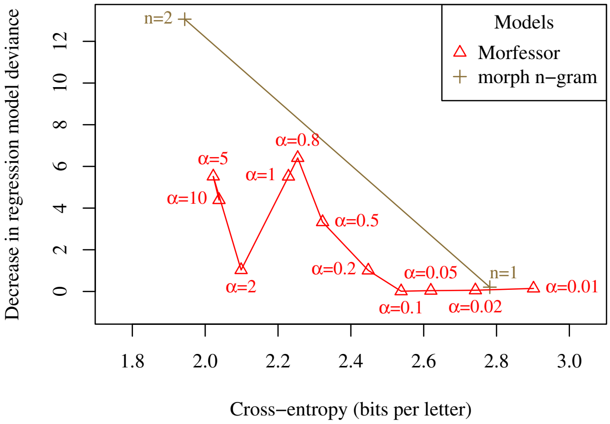

Fig. 6. Cognitive prediction accuracy (decrease in deviance) versus text prediction accuracy (cross-entropy) of the language models; all control predictors included. The improvement over the baseline regression model is statistically significant ($p_{\text{anova}} \leq .05$) for the morph bigram model and Morfessor Baseline models at $5 \leq \alpha \leq 10$ and $0.8 \leq \alpha \leq 1.0$.

section, we calculated segmentation boundary precision, recall, and F-measure, and categorized the precision and recall disparities based on the surrounding morph types. Table 4 shows the different types of disparities in the segmentations of Morfessor Baseline models at $\alpha = 0.01$ (closest to linguistic morpheme-level segmentation according to F-measure) and $\alpha = 0.8$ (best cognitive prediction accuracy). For example, when $\alpha = 0.01$, Morfessor has placed a segmentation boundary to 381 of the possible 1,910 boundaries within stems (19.9% of the maximum), whereas at $\alpha = 0.8$, there are only 51 boundaries within stems (2.7% of the maximum). The precision values of the two models are 0.433 and 0.678, recall values 0.511 and 0.283, and F-measures 0.469 and 0.399, respectively. Examples of the segmentations produced by the models are shown in Table 5.

The numbers of recall disparities for the full range of Morfessor Baseline models are shown in Fig. 7. Most of the disparity types are increasing quite consistently. However, a large part of the boundaries between two derivational suffixes are missed already with low values of α. In contrast, some clitics cease to be split only when $\alpha \geq 3$.

4. Discussion

We investigated human morphological processing by using an MDL-based computational model Morfessor Baseline (Creutz & Lagus, 2002, 2005a,b, 2007) that works on the principle of optimization. We asked what this kind of a model can tell us about optimal units of representation and the cognitive architecture within the mental lexicon. Morfessor utilizes rather simple learning principles and is unsupervised, that is, it creates a morphological lexicon based solely on observing individual words and is thus capable of learning without supervision. We compared models that utilize pre-segmented linguistic

Table 3
Contribution of word-specific variables to the regression model when a language model is used as a control predictor

Language Model (control)	Predictor	β	AIC	p_{anova}
Word unigram	Num. of letters	0.2573	−6695.6	.0000
Morfessor α = 0.8	Num. of letters	0.2043	−6805.3	.0000
Morph bigram	Num. of letters	0.2328	−6791.3	.0000
Word unigram	Num. of morphs	0.1346	−6599.1	.0000
Morfessor α = 0.8	Num. of morphs	0.0774	−6747.2	.0007
Morph bigram	Num. of morphs	0.0746	−6702.3	.0028
Word unigram	Surface frequency	−0.0241	−6564.3	.1216
Morfessor α = 0.8	Surface frequency	−0.0204	−6752.7	.0000
Morph bigram	Surface frequency	−0.0259	−6719.5	.0000
Word unigram	Lemma frequency	−0.0058	−6565.7	.0506
Morfessor α = 0.8	Lemma frequency	−0.0036	−6737.4	.1974
Morph bigram	Lemma frequency	−0.0112	−6710.3	.0000
Word unigram	Morph. family size	−0.0056	−6564.4	.1117
Morfessor α = 0.8	Morph. family size	−0.0047	−6737.8	.1489
Morph bigram	Morph. family size	−0.0095	−6701.6	.0039
Word unigram	Infl. entropy	0.0063	−6564.6	.1016
Morfessor α = 0.8	Infl. entropy	0.0030	−6736.4	.4107
Morph bigram	Infl. entropy	0.0110	−6702.8	.0021
Word unigram	Relative entropy	0.0080	−6569.2	.0070
Morfessor α = 0.8	Relative entropy	0.0061	−6740.5	.0289
Morph bigram	Relative entropy	0.0117	−6711.4	.0000
Word unigram	NDR	0.0176	−6608.1	.0000
Morfessor α = 0.8	NDR	0.0122	−6755.9	.0000
Morph bigram	NDR	0.0137	−6714.6	.0000

Note. With morph 2-gram as a control predictor, all variables yield significant contributions. With word 1-gram as a control predictor, surface frequency, morphological family size, and inflectional entropy do not provide significant improvements. With Morfessor Baseline (α = 0.8) as a control predictor, lemma frequency, morphological family size, and inflective entropy do not provide significant improvements.

morphs (morph *n*-gram models) to the different variants of Morfessor which can also yield other kinds of units (e.g., longer than linguistic morphs). As a measure of whole word frequency we studied the performance of the word unigram model along with the morph-based models. The aim was to see what kind of model structure corresponds best to human word recognition costs in processing multimorphemic Finnish words. We thus compared the performance of these statistical models of morphology in predicting RTs in a visual lexical decision task. Throughout the results, we took into account models' cross-entropy, shown to be closely linked to their cognitive prediction accuracy (e.g., Frank, 2009), as we are interested in how well other aspects of the models apart from cross-entropy perform in predicting RTs.

The results generally show that the best RT predictions were reached by the unsupervised Morfessor and specifically its instance at α = 0.8, which decomposes some words at (some of) their morpheme boundaries and keeps others unsegmented. It performed

Table 4
Precision and recall disparities for segmentations of Morfessor Baseline α = 0.01 and α = 0.8

		α = 0.01		α = 0.8	
	max.	#	%	#	%
Precision disparities					
STEM	1,910	381	19.9	51	2.7
DERIV	285	25	8.8	16	5.6
INFL	453	21	4.6	19	4.2
CLITIC	71	1	1.4	0	0
Total	2,719	428	14.7	86	3.2
Recall disparities					
STEM+DERIV	121	55	45.5	94	77.7
STEM+INFL	196	61	31.1	129	65.8
STEM+CLITIC	8	0	0	0	0
DERIV+DERIV	21	13	61.9	20	95.2
DERIV+INFL	101	72	71.3	85	84.2
DERIV+CLITIC	3	0	0	0	0
INFL+INFL	169	111	65.7	131	77.5
INFL+CLITIC	21	1	4.8	0	0
Total	640	313	48.9	459	71.7

Note. Columns show the type of disparity, maximum number of disparities for the type (max.), number of found disparities (#), and ratio of the found disparities to the maximum disparities (%). A precision disparity means that the method has inserted a boundary within a linguistic morph of a certain type; a recall disparity means that the method has not inserted a boundary between two linguistic morphs.

better than the supervised models that are strictly based on linguistic morphs. This finding suggests that linguistic morphs may not always be the primary processing units within the mental lexicon. On the other hand, the whole-word based word unigram model did not perform as well as Morfessor, either.

Overall, the results confirmed that self-information of a word (see, e.g., Boston et al., 2008; Frank, 2009; Levy, 2008), as determined by a statistical language model, correlates strongly with human word recognition costs in a lexical decision task that includes morphologically complex Finnish nouns (Table 1). The psycholinguistic control variables mostly showed the typically observed correlations with the RTs; for example, both lemma frequency and morphological family size showed a significant negative correlation (e.g., Bertram et al., 2000; Taft, 1979), whereas the relative entropy measure correlated positively with the RTs (e.g., Milin, Kuperman et al., 2009). Inflectional entropy also correlated positively with the RTs. This is in line with the assumption that lexical units with higher information load are more costly to retrieve, although studies have also reported facilitatory effects for this variable in word recognition (e.g., Baayen et al., 2006). We also observed that cross-entropy and the model's ability to predict recognition times displayed a correlation: High text prediction accuracy tends to imply high cognitive prediction accuracy (Fig. 3). Interestingly, however, some of the language models predicted RTs better than was to be expected on the basis of their cross-entropies.

Table 5
Examples of stimulus words segmented according to the linguistic analyzer and Morfessor Baseline

Word	Linguistic Segmentation	Baseline α = 0.01	Baseline α = 0.8
haastajaksi	haasta $_{\text{V}}$ ja $_{\text{+ DV-JA}}$ ksi $_{\text{+TRA}}$	haasta ja ksi	haastaja ksi
as a challenger	*challenge [-r] [transitive]*		
julkaisuineen	julkais $_{\text{V}}$ u $_{\text{+DV-U}}$ ine $_{\text{+CMT}}$ en $_{\text{+3SGPL}}$	julkaisu ineen	julkaisu ineen
with her publications	*publish [verb to noun] [comitative] [her/his]*		
kattilaan	kattila $_{\text{N}}$ an $_{\text{+ILL}}$	kat tila an	kattilaan
into a kettle	*kettle [il lative]*		
maksujaankaan	maksu $_{\text{N}}$ j $_{\text{+PL}}$ a $_{\text{+PTV}}$ an $_{\text{+3SGPL}}$ kaan $_{\text{CLI}}$	maksu ja an kaan	maksuja an kaan
her payments either	*payment [plural] [partitive] [her/his] [either]*		
monologissaan	monologi $_{\text{N}}$ ssa $_{\text{+INE}}$ an $_{\text{+3SGPL}}$	mon ologi ssa an	monologi ssaan
in her monologue	*monologue [inessive] [her/his]*		
ohjaajana	ohjaa $_{\text{V}}$ ja $_{\text{+DV-JA}}$ na $_{\text{+ESS}}$	ohjaaja na	ohjaajana
as the instructor	*instruct [-or] [essive]*		
peruutuksestasi	peruut $_{\text{V}}$ ukse $_{\text{+DV-US}}$ sta $_{\text{+ELA}}$ si $_{\text{+2SG}}$	peruu tuksesta si	peruutuksesta si
about your cancel lation	*cancel [-lation] [elative] [your]*		
porojen	poro $_{\text{N}}$ j $_{\text{+PL}}$ en $_{\text{+GEN}}$	poro jen	porojen
of reindeers	*reindeer [plural] [genitive]*		
vaikeuksiakin	vaike $_{\text{A}}$ uks $_{\text{+DA-US}}$ i $_{\text{+PL}}$ a $_{\text{+PTV}}$ kin $_{\text{CLI}}$	vaike uksia kin	vaikeuksia kin
also difficulties	*difficult [-y] [plural] [also]*		
ystävällenne	ystävä $_{\text{N}}$ lle $_{\text{+ALL}}$ nne $_{\text{+2PL}}$	ystävä lle nne	ystävälle nne
to your friend	*friend [al lative] [your]*		

Note. In the linguistic segmentation, subscripts mark the morph categories: A, N, and V refer to adjective, noun, and verb stems, respectively, and DA-, DN-, and DV- to their derivational suffixes. Inflectional suffixes start with a plus sign. Clitics are marked by CLI.

4.1. The best-performing model segments at some but not all morpheme boundaries

When comparing the performance of the different types of morphological models, an instance of Morfessor (at α = 0.8) performed the best in predicting RTs when no variables apart from word presentation order were included in the analysis. The supervised morph unigram model, which bases its analysis on linguistic morphs, has a similar structure as Morfessor as it assumes that morphemes occur independently of one another. The performance of this supervised model was not as high as that of Morfessor, suggesting that linguistic morphemes are too short to give a good estimate of self-information of the whole word. Inaccurate self-information estimates are also indicated by the higher cross-entropy of the supervised model. Morfessor optimizes the likelihood of the training data as part of its cost function and thus also reaches a lower cross-entropy. Morfessor, however, also outperformed the supervised model that has a more favorable cross-entropy value, that is, the morph bigram model which takes into account the context in which individual morphemes occur. An MDL-based statistical model trained in an unsupervised manner was thus able to predict RTs more accurately than this supervised implementation. While Morfessor often produces segmentations that correspond to linguistic morphemes, it offers cognitively more accurate predictions than models solely based on linguistic morphemes.

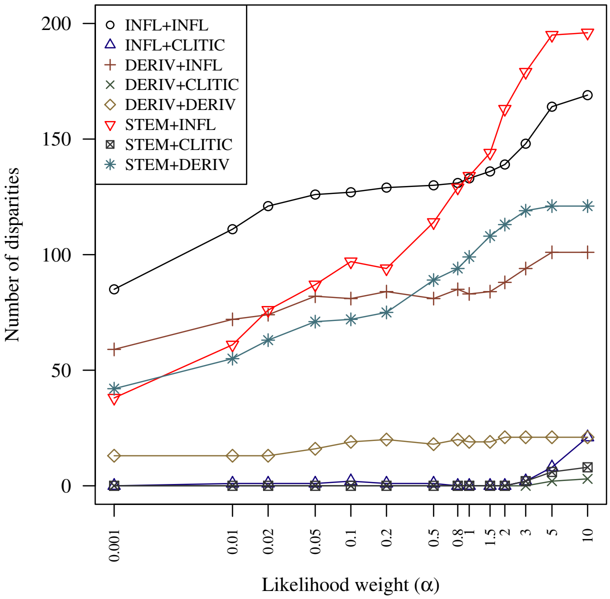

Fig. 7. Boundary recall disparities for Morfessor segmentations with varying likelihood weight parameter. A recall disparity is a segmentation boundary in the linguistic morphological segmentation that is not predicted by the model. The maximum number of disparities is shown by the model instance at $\alpha = 10$. Most of the disparity types increase consistently. However, a large part of the boundaries between two derivational suffixes are missed already at low values of α. In contrast, some clitics cease to be split only when $\alpha \geq 3$.

To further investigate optimal units of lexical representation, we were able to manipulate the same model's emphasis on decomposition versus full-form storage by assessing the performance of Morfessor Baseline for different values of the likelihood weight parameter α. Low values of α in Morfessor are associated with more extensive morphological segmentation and high values with dominating full-form storage.

The model instance at $\alpha = 0.01$ produced morphs that resembled linguistic segmentations the closest (Fig. 2). This model instance did not show particularly high cognitive prediction accuracy. This result is in line with the observation above that the supervised models based on linguistic morphemes did not fare very well in their present RT predictions.

With a high value of α, all words are in practice stored in the lexicon. Increasing the value of α from 0.8 to 10 did not lead to an improved cognitive prediction accuracy, although the model instance at $\alpha = 10$ had a lower, that is, better, cross-entropy value than the one at $\alpha = 0.8$. The same was true for the word unigram model which has the best cross-entropy in the present set of models. This is a different pattern of results than that of Frank (2009), who found that in sentence processing cross-entropy had a monotonous relation to the accuracy in predicting RTs within each model type. Although the models based on full forms had good text prediction accuracy, they were not the best predictors of word recognition times in the present study.

Our results do not rule out the possibility that lexical items corresponding to human processing could be longer than those found at $\alpha = 0.8$. Because of the limitation observed at $\alpha = 2.0$ (described in Section 3.1), cognitive prediction accuracies for Morfessor instances that select a full form representation for a large proportion of the words may appear lower than could be reached without the limitation. Thus, the value 0.8 can be considered as a lower boundary for the optimal α. Moreover, the optimal value must be clearly under 10, as the limitation does not apply there, and the performance at $\alpha = 10$ is still worse than that at $\alpha = 0.8$.

When investigating the segmentations produced by the highest-performing Morfessor instance (at $\alpha = 0.8$) in the current set of words, we found that it left all clitic particles distinct (see Tables 4 and 5; e.g., the word *vaikeuksiakin*). Moreover, it did not segment words at the majority of boundaries that were followed by a derivational suffix (see Table 5; e.g., the word *ohjaajana*). Several studies (see, e.g., Bozic & Marslen-Wilson, 2010; Laudanna, Badecker, & Caramazza, 1992; Niemi et al., 1994) suggest that derivations, which are often semantically less transparent and less productive than, for example, inflected words, are processed as holistic units via full-form representations. Inflected words are often assumed to be fully decomposed, although full-form representations have also been proposed for high-frequency word forms (see, e.g., Alegre & Gordon, 1999; Baayen et al., 1997; Lehtonen & Laine, 2003; Soveri et al., 2007). Here, however, the best Morfessor model variant left two-thirds of bimorphemic stem + inflectional suffix combinations unsegmented (Table 5; e.g., the word *kattilaan*), and the same was true with the majority of derivation + inflection boundaries (see Table 5; e.g., the words *peruutuksestasi* and *ohjaajana*). Thus, a model which also allows full-form recognition for many complex words, both derived and inflected ones, performed better than a model which segments all complex words exhaustively into their morphemic constituents. While Morfessor at $\alpha = 0.8$ did not segment all linguistically determined morpheme boundaries, it should be noted that it sometimes produced segmentations that were located within the morpheme, and thus in implausible positions. However, this took place only in 3.2% of the morpheme boundaries in the stimulus words (Table 4).

In the Finnish language, many nouns have multiple stem allomorphs and undergo stem changes when inflected (e.g., kenkä; kengä+n). Morfessor Baseline does not specifically model allomorphic variations, so the possible allomorphs need to be stored separately in the model's lexicon. Morfessor segments some of these word forms into their stem allomorphs, depending on how frequently the allomorph occurs in different words. Behavioral evidence from Finnish adults in fact shows that different allomorphs also have their own lexical representations (Järvikivi & Niemi, 2002; Niemi et al., 1994).

4.2. The best prediction ability is obtained with both decomposition and full-form measures

The best cognitive prediction accuracy was found for Morfessor at $\alpha = 0.8$ in an analysis which did not include any control predictors apart from word order number in the regression model. This instance of the MDL-based Morfessor thus seemed to capture relevant aspects of the human word recognition process. With regard to its likelihood weight α, this model instance was in the middle ground in the range between emphasizing full-forms versus decomposition, that is, it segments words at some morpheme boundaries, but it also leaves many of the boundaries unsegmented.

When the instance of Morfessor at $\alpha = 0.8$ was included in the regression analysis together with the word unigram model, Morfessor could still improve the predictions. Thus, a regression model that included measures that allow both decomposition and whole-word processing was better able to predict the processing costs of human word recognition than a model that included only one type of measure. This suggests that the frequency of the whole word also plays an independent role in word recognition.

As the different word-related control variables are likely to explain partly similar variance as our language models (but differently for each specific model), our primary analysis was the one without any control predictors in order to study the optimal processing units of the mental lexicon without the influence of these variables. However, we also studied how each of the models relate to known psycholinguistic variables such as lemma frequency, word length, and morphological family size, as well as the NDR model (Baayen et al., 2011). That is, to what extent do these psycholinguistic variables capture the same variance in the RTs as the statistical language models. When the word-specific variables were included as control predictors in the regression model, Morfessor Baseline at $\alpha = 0.8$ and the morph bigram model further improved the predictions. This indicates that they add something relevant to the known word-specific psycholinguistic variables in explaining variance in the RTs. The analysis which included different language models as control predictors for each word-specific variable (see Table 2) showed that Morfessor at $\alpha = 0.8$ was able to capture to a large extent similar variance as morphological family size, lemma frequency, and inflectional entropy but not that of surface frequency, word length (in letters or morphs), or NDR. Thus, Morfessor explains largely morphological aspects of word recognition. The morph bigram model also clearly showed an independent effect, likely because it predicts upcoming morphs based on the previous ones, which is an aspect of multimorphemic word processing not directly captured by the included lexical variables.

Both Morfessor and a whole-word model provided independent contributions to RT predictions within the same regression model (Tables B4 and B6 in Appendix S2). Dual-route models of morphological processing (e.g., Baayen et al., 1997; Frauenfelder & Schreuder, 1992; Schreuder & Baayen, 1995) assume that both decomposed and full-form representations are simultaneously activated and these processing “routes” thus work in parallel. On the basis of data on compound processing, Kuperman et al. (2009) have sketched a multiple-route model of morphological processing that would allow access to full forms, morphological constituents, and morphological families at different times and to a different extent. According to Kuperman et al. (2009), readers take advantage of multiple sources of information in a parallel and interactive way. Such a model could also explain the present findings. On the other hand, it has been proposed that measures of decomposition and full-form processing may reflect different stages of word recognition (e.g., Fruchter & Marantz, 2015;

Taft, 2004): Decomposition at the visual word form level may be more sensitive to measures of decomposition (see also Rastle & Davis, 2008), whereas a later recombination stage in which the meaning of decomposed morphemes is integrated would be sensitive to measures of the whole word, that is, combination of the morphemes. As the present study used simple RTs which provide an end-point measure of the entire recognition process, either or both of these alternatives about the word recognition process could be correct. Time-sensitive neuroimaging may provide opportunities to specifically target different levels of morphological processing (see, e.g., Lehtonen et al., 2007, 2006; Vartiainen et al., 2009). Future work should determine whether the predictions of the models tested here might be specific to particular processing levels.

The present study investigated lexical processing in adult native speakers that is looked at the processes in an established language system and therefore speaks to the issues of language learning only indirectly. Nevertheless, the results show that an unsupervised model, using general statistical learning principles corresponds better to human word recognition than a model utilizing only linguistically structured units. In fact, there are similarities between the learning process of Morfessor and how learning of morphological regularities has been suggested to take place in humans as well (e.g., Schreuder & Baayen, 1995): The process may start from forming initial whole-word representations of the observed input, proceeding to discovering structural regularities, and forming morpheme-based representations. With increased exposure to commonly occurring morpheme combinations, it will become economical to store such chunks as full forms again (for evidence of storage of inflected forms from Finnish children, see, e.g., Räsänen, Ambridge, & Pine, 2016). According to the present results, the adult system seems to code some words as full forms and process others as decomposed parts, and, for some words at least, utilize both kinds of representations in their processing. While commonly occurring morpheme combinations may develop full-form representations, it is unlikely that morpheme-based representations would altogether vanish in this process. Such representations are needed when encountering novel words including these morphemes or words in which this morpheme is combined with an unusual affix or compound constituent.

Apart from the dual-route model framework, it is interesting to consider particular alternative accounts that might be used to describe processing of complex words. The present study focused on statistical models of morphology which give self-information estimates and which assume that morphemes may play a role in the architecture of the mental lexicon. This choice enabled us to investigate the optimal balance between decomposition and full-form recognition in the human mental lexicon. At the same time, this focus leaves out implementations based on other principles, such as the NDR (Baayen et al., 2011) which maps orthographic units directly to meanings via a discriminative learning mechanism without a morphological (or lexical) level. However, we included the NDR model as a control variable and found that both Morfessor and NDR contribute independently to the RTs (see Table 3) and can thus be interpreted to describe different aspects of word recognition costs. It appears that together they provide better prediction ability than either one does alone. An intriguing possibility would be to allow the units provided by Morfessor to serve as input cues to the NDR model (instead of the predetermined letter bigrams or trigrams in Baayen et al., 2011; or triphones in Baayen et al., 2016). Such a combination could provide more accurate predictions than either model alone and capture relevant processing aspects at different levels of word recognition, from visual processing to lexical units and further to semantics.

4.3. Conclusions

The present results show that a computational model that works in an unsupervised manner, using the MDL optimization principle performs well in predicting recognition times of morphologically complex words. The best-performing Morfessor instance was one that decomposes words at some morpheme boundaries and keeps other boundaries unsegmented. Unsegmented boundaries were found especially in words containing derivational suffixes but also for a large part of words with inflectional suffixes. This kind of implementation corresponded better to human word recognition times than supervised models based solely on linguistic morphemes or those that only included whole word forms. However, an even better prediction accuracy was provided by a combination of a Morfessor model and a word unigram model based on full forms. These results support cognitive models that assume that both kinds of representations, decomposed and full form representations, are utilized in order to optimally process and store complex words within the mental lexicon.

Acknowledgments

The authors were supported by the Academy of Finland (http://www.aka.fi) via LASTU Programme grants 256887 and 259934 and personal grants 288880 to ML, 287474 to AH, and 255349, 256459, and 283071 to RS. We thank Mikko Kurimo for the discussion and comments on the manuscript and Matti Varjokallio for his help on processing the data sets.

References

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

Ambridge, B., & Lieven, E. (2015). A constructivist account of child language acquisition. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 478–510). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118346136.ch22

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

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

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512. https://doi.org/10.1016/j.jml.2006.03.008

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

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128. https://doi.org/10.1080/23273798.2015.1065336

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

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

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

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12. https://doi.org/10.16910/jemr.2.1.1

Bozic, M., & Marslen-Wilson, W. (2010). Neurocognitive contexts for morphological complexity: Dissociating inflection and derivation. *Language and Linguistics Compass*, *4*, 1063–1073. https://doi.org/10.1111/j.1749-818X.2010.00254.x

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language production* (pp. 257–294). London: Academic Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., & Stolcke, A. (2007). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, *5*(1), 1–29. https://doi.org/10.1145/1322391.1322394

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In M. Maxwell (Ed.), *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 21–30). Shroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1118647.1118650

Creutz, M., & Lagus, K. (2005a). Inducing the morphological lexicon of a natural language from unannotated text. In T. Honkela, V. Könönen, M. Pöllä, & O. Simula (Eds.), *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (pp. 106–113). Espoo, Finland: Helsinki University of Technology, Laboratory of Computer and Information Science. Available at http://research.ics.aalto.fi/events/AKRR05/papers/akrr05creutz.pdf

Creutz, M., & Lagus, K. (2005b). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0 (tech. rep. No. A81). Publications in Computer and Information Science, Helsinki University of Technology. Available at http://www.cis.hut.fi/mcreutz/papers/Creutz05tr.pdf

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, *4*(1), 1–34. https://doi.org/10.1145/1187415.1187418

Creutz, M., & Lindén, K. (2004). Morpheme segmentation gold standards for Finnish and English (tech. rep. No. A77). Publications in Computer and Information Science, Helsinki University of Technology. Available at http://www.cis.hut.fi/mcreutz/papers/Creutz04tr.pdf

Diependaele, K., Sandra, D., & Grainger, J. (2005). Masked cross-modal morphological priming: Unravelling morpho-orthographic and morpho-semantic influences in early word recognition. *Language and Cognitive Processes*, *20*(1–2), 75–114. https://doi.org/10.1080/01690960444000197

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain & Language*, *129*, 14–23. https://doi.org/10.1016/j.bandl.2013.11.004

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In R. Levy & D. Reitter (Eds.), *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)* (pp. 61–69). Montreal, Canada: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/W/W12/W12-1706.pdf

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the annual meeting of the Cognitive Science Society* (pp. 1139–1144). Austin, TX: Cognitive Science Society. Available at http://escholarship.org/uc/item/02v5m1hf.pdf

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*, 475–494. https://doi.org/10.1111/tops.12025

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. https://doi.org/10.1177/0956797611409589

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. Booij & van Merle J. (Eds.), *Yearbook of morphology 1991* (pp. 165–183). Dordrecht, the Netherlands: Kluwer. https://doi.org/10.1007/978-94-011-2516-1_10

Fruchter, J., & Marantz, A. (2015). Decomposition, lookup, and recombintion: MEG evidence for the full decomposition model of complex visual word recognition. *Brain and Language*, *143*, 81–96. https://doi.org/10.1016/j.bandl.2015.03.001

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153–189. https://doi.org/10.1162/089120101750300490

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 459–466). Cambridge, MA: MIT Press. Available at http://papers.nips.cc/paper/2941-interpolating-between-types-and-tokens-by-estimating-power-law-generators.pdf

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, *12*, 2335–2382.

Grasemann, U., Kiran, S., Sandberg, C., & Miikkulainen, R. (2011). Impairment and rehabilitation in bilingual aphasia: A SOM-based model. In J. Laaksonen & T. Honkela (Eds.), *Proceedings of WSOM11, 8th workshop on self-organizing maps* (Vol. 6731, pp. 207–217). Lecture Notes in Computer Science. Espoo, Finland: Springer Verlag. https://doi.org/10.1007/978-3-642-21566-7_21

Hafer, M. A., & Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, *10*(11–12), 371–385. https://doi.org/10.1016/0020-0271(74)90044-8

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In R. Levy & R. Reitter (Eds.), *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)* (p. 8). Pittsburgh, PA:: Association for Computational Linguistics. Available at http://aclweb.org/anthology/N/N01/N01-1021.pdf

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*(2), 190–222. https://doi.org/10.2307/411036

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, *9*(7), 342–348. https://doi.org/10.1016/j.tics.2005.04.002

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., & Pylkkönen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, *20*(4), 515–541. https://doi.org/10.1016/j.csl.2005.07.002

Hsu, A. S., & Chater, N. (2011). The logical problems of language acquisition: A probabilistic perspective. *Cognitive Science*, *34*, 972–1016. https://doi.org/10.1111/j.1551-6709.2010.01117.x

Hyönä, J., Laine, M., & Niemi, J. (1995). Effects of a word's morphological complexity on readers' eye fixation patterns. In J. Findlay, R. Kentridge, & R. Walker (Eds.), *Eye movement research: Mechanisms,* *processes and applications* (pp. 445–452). Amsterdam: North-Holland. https://doi.org/10.1016/s0926-907x(05)80037-6

Järvikivi, J., & Niemi, J. (2002). Form-based representation in the mental lexicon: Priming (with) bound stem allomorphs in finnish. *Brain and Language*, *81*(1), 412–423. https://doi.org/10.1006/brln.2001.2534

Kohonen, O., Virpioja, S., & Lagus, K. (2010). Semi-supervised learning of concatenative morphology. In J. Heinz, L. Cahill, & R. Wicentowski (Eds.), *Proceedings of the 11th meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (pp. 78–86). Uppsala, Sweden: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/W/W10/W10-2210.pdf

Koskenniemi, K. (1983). Two-level morphology: a general computational model for word-form recognition and production (Doctoral dissertation, University of Helsinki, Helsinki, Finland). Available at http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf. Accessed October 17, 2017

Kostić, A. (1991). Informational approach to processing inflected morphology: Standard data reconsidered. *Psychological Research*, *53*(1), 62–70. https://doi.org/10.1007/BF00867333

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*, 83–97. https://doi.org/10.1016/j.jml.2009.10.001

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

Kurimo, M., Creutz, M., & Varjokallio, M. (2008). Morpho Challenge evaluation using a linguistic gold standard. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, V. Petras, & D. Santos (Eds.), *Advances in multilingual and multimodal information retrieval*, 8th workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised selected papers (Vol. 5152, pp. 864–873). Berlin: Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-540-85760-0_111

Lagus, K., Kohonen, O., & Virpioja, S. (2009). Towards unsupervised learning of constructions from text. In M. Sahlgren & O. Knutsson (Eds.), *Proceedings of the workshop on extracting and using constructions in NLP of the 17th Nordic conference on computational linguistics (NODALIDA). SICS Technical Report T2009:10*. Odense, Denmark: Swedish Institute of Computer Science. Available at http://soda.swedish-ict.se/3627/1/SICS-T–2009-10–SE.pdf

Laine, M., Vainio, S., & Hyönä, J. (1999). Lexical access routes to nouns in a morphologically rich language. *Journal of Memory and Language*, *40*, 109–135. https://doi.org/10.1006/jmla.1998.2615

Laine, M., & Virtanen, P. (1996). *Wordmill lexical search program*. Centre for Cognitive Neuroscience: University of Turku.

Laudanna, A., Badecker, W., & Caramazza, A. (1992). Processing inflectional and derivational morphology. *Journal of Memory and Language*, *31*, 333–348. https://doi.org/10.1016/0749-596X(92)90017-R

Lehtonen, M., Cunillera, T., Rodríguez-Fornells, A., Hultén, A., Tuomainen, J., & Laine, M. (2007). Recognition of morphologically complex words in Finnish: Evidence from event-related potentials. *Brain Research*, *1148*, 123–137. https://doi.org/10.1016/j.brainres.2007.02.026

Lehtonen, M., & Laine, M. (2003). How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and Cognition*, *6*, 213–225. https://doi.org/10.1017/s1366728903001147

Lehtonen, M., Vorobyev, V. A., Hugdahl, K., Tuokkola, T., & Laine, M. (2006). Neural correlates of morphological decomposition in a morphologically rich language: An fMRI study. *Brain and Language*, *98*(2), 182–193. https://doi.org/10.1016/j.bandl.2006.04.011.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics*, *1*, 333–353. https://doi.org/10.1146/annurev-linguist-030514-125236

Milin, P., Đurđević, D. F., & del Prado Martín, F. M. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, *60*(1), 50–64. https://doi.org/10.1016/j.jml.2008.08.007

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 214–252). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199547548.003.0010

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. https://doi.org/10.1037/0278-7393.30.6.1271

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. https://doi.org/10.1016/j.cognition.2003.10.015

Niemi, J., Laine, M., & Tuominen, J. (1994). Cognitive morphology in Finnish: Foundations of a new model. *Language and Cognitive Processes*, *9*, 423–446. https://doi.org/10.1080/01690969408402126

Norris, D. (2006). April). The Bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, *113*(2), 327–357. https://doi.org/10.1037/0033-295X.113.2.327

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462. https://doi.org/10.1016/S1364-6613(98)01241-8

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*(6), 1199–1241. https://doi.org/10.1162/08997660152002834

Quasthoff, U., Richter, M., & Biemann, C. (2006). Corpus portal for search in monolingual corpora. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the fifth international conference on language resources and evaluation, LREC 2006* (pp. 1799–1802). Genoa, Italy: European Language Resources Association. Available at http://www.lrec-conf.org/proceedings/lrec2006/pdf/641_pdf.pdf

Räsänen, S. H. M., Ambridge, B., & Pine, J. M. (2016). An elicited-production study of inflectional verb morphology in child Finnish. *Cognitive Science*, *40*(7), 1704–1738. https://doi.org/10.1111/cogs.12305

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*(7–8), 942–971. https://doi.org/10.1080/01690960802069730

Rescorla, R. (2007). Rescorla-Wagner model. *Scholarpedia*, *2*(3), 2237. https://doi.org/10.4249/scholarpedia.2237

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

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Hackensack, NJ: World Scientific.

Rueckl, J. G. (2010). Connectionism and the role of morphology in visual word recognition. *The Mental Lexicon*, *5*(3), https://doi.org/10.1075/ml.5.3.07rue

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 131–154). Hillsdale, NJ: Lawrence Erlbaum.

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

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Soveri, A., Lehtonen, M., & Laine, M. (2007). Word frequency and morphological processing in Finnish revisited. *The Mental Lexicon*, *2*, 359–385. https://doi.org/10.1075/ml.2.3.04sov

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

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology Section A*, *57*(4), 745–765. https://doi.org/10.1080/02724980343000477

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

Tiedemann, J. (2009). News from OPUS — A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent advances in natural language processing* (vol V) (pp. 237–248). Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.309.19tie

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Vartiainen, J., Aggujaro, S., Lehtonen, M., Hultén, A., Laine, M., & Salmelin, R. (2009). Neural dynamics of reading morphologically complex words. *NeuroImage*, *47*(4), 2064–2072. https://doi.org/10.1016/j.neuroimage.2009.06.002

Virpioja, S., Kohonen, O., & Lagus, K. (2011). Evaluating the effect of word frequencies in a probabilistic generative model of morphology. In B. S. Pedersen, G. Nešpore, & I. Skadiņa (Eds.), *Proceedings of the 18th Nordic conference of computational linguistics (NODALIDA 2011)* (Vol. 11, pp. 230–237). NEALT Proceedings Series. Riga, Latvia: Northern European Association for Language Technology. Available at http://hdl.handle.net/10062/17313

Virpioja, S., Lehtonen, M., Hultén, A., Salmelin, R., & Lagus, K. (2011). Predicting reaction times in word recognition by unsupervised learning of morphology. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), *Artificial neural networks and machine learning — ICANN 2011* (Vol. 6791, pp. 275–282). Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/978-3-642-21735-7_34

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1189–1198). Stroudsburg, PA: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/P10-1121.pdf

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456. https://doi.org/10.1016/j.tics.2004.08.006

Zipf, G. K. (1932). *Selective studies and the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Implementation of the statistical models and predictors.

Appendix S2. Details of the regression models.