Predicting Reaction Times in Word Recognition by Unsupervised Learning of Morphology

Sami Virpioja[1], Minna Lehtonen[2,3,4], Annika Hultén[3,4],
Riitta Salmelin[3], and Krista Lagus[1]

[1] Department of Information and Computer Science,
Aalto University School of Science
[2] Cognitive Brain Research Unit, Cognitive Science,
Institute of Behavioural Sciences, University of Helsinki
[3] Brain Research Unit, Low Temperature Laboratory,
Aalto University School of Science
[4] Department of Psychology and Logopedics, Åbo Akademi University

Abstract. A central question in the study of the mental lexicon is how morphologically complex words are processed. We consider this question from the viewpoint of statistical models of morphology. As an indicator of the mental processing cost in the brain, we use reaction times to words in a visual lexical decision task on Finnish nouns. Statistical correlation between a model and reaction times is employed as a goodness measure of the model. In particular, we study Morfessor, an unsupervised method for learning concatenative morphology. The results for a set of inflected and monomorphemic Finnish nouns reveal that the probabilities given by Morfessor, especially the Categories-MAP version, show considerably higher correlations to the reaction times than simple word statistics such as frequency, morphological family size, or length. These correlations are also higher than when any individual test subject is viewed as a model.

1 Introduction

The processing of morphologically complex words is a central question in the study of the mental lexicon. Theoretical models have been put forward that suggest that morphologically complex words are recognized either through full-form representations [3], full decomposition (e.g. [17]) or a combination of the two (e.g. [11]). For example, Finnish words can be combined of several morphemes, and one single noun can, in principle, attain up to 2000 different forms [7]. Having separate neural representations for each of these forms would seem unnecessarily demanding compared to a process where words would be analyzed based on their compound morphemes. In behavioral word recognition tasks, a processing cost (i.e., long reaction times and high error rates) has been robustly associated with inflected Finnish nouns in comparison to matched monomorphemic nouns [11,10]. This has been taken as evidence for the existence of morphological decomposition for most Finnish inflected words, with the possible exception of very high frequency inflected nouns [15].

T. Honkela et al. (Eds.): ICANN 2011, Part I, LNCS 6791, pp. 275–282, 2011.

Statistical models of language learning would be attractive both conceptually and because they yield quantitative predictions that may be tested against measured values of performance and, eventually, of brain activation. In this first feasibility test, we use reaction times as a proxy, providing an indirect measure of the underlying mental processing. In previous studies, several factors, including the cumulative base frequency (i.e., the summative frequency of all the inflectional variants of a single stem, [16]), surface frequency (i.e., whole form frequency, [1]), and morphological family size (i.e., the number of derivations and compounds where the noun occurs as a constituent, [2]), have been found to affect the recognition times of morphologically complex words. However, we do not know of any previous work that would use statistical models of morphology as models of the reaction times. In the proposed evaluation setting, we examine how well they predict the average reaction times for individual inflected and monomorphemic words in a word recognition task. As a particular morphological model we examine an unsupervised method for word segmentation, Morfessor, that induces a compact lexicon of morphs from unannotated text data.

2 Experimental Setup

Our experimental setup can be summarized as follows: (1) *Data recording:* Measurement data from humans is obtained, namely reaction times recorded on test subjects in a lexical decision task with inflected and monomorphemic words. (2) *Model estimation:* Using training data of varying size and type, we estimate statistical models of morphology that can be used to predict the recognition times of words. In addition, we collect such statistics of the words that are known to affect the reaction times. (3) *Model evaluation:* We calculate linear correlation between model predictions and the average reaction times of the test subjects. A good model is one which produces costs that have high correlation to the reaction times. Also any of the human test subjects can be viewed as a model, and their reaction times thus correlated with those of the rest of the subjects.

2.1 Reaction Time Data and Model Evaluation

We use the reaction time data reported in [9]. Sixteen Finnish-speaking university students participated in the experiment. The task was to decide as quickly and accurately as possible whether the letter string appearing on the screen was a real Finnish word or not, and to press a corresponding button. The stimuli consisted of 320 real Finnish nouns and 320 pseudowords. The words were taken from an unpublished Turun Sanomat newspaper corpus of 22.7 million word tokens and divided into four groups of 80 words according to their frequency in the corpus (high or low) and morphological structure (monomorphemic or inflected). There were four kinds of pseudowords (monomorphemic, real stem with pseudosuffix, pseudostem with real suffix, and incorrect combination of real stem and suffix) and their lengths and bigram frequencies (i.e., the average frequency of letter bigrams in the word) were similar to the real words.

As preprocessing, we exclude all incorrect responses and reaction times of three standard deviations longer or shorter than the individual's mean. For the remaining data, we take the logarithm of the reaction times, normalize them to zero mean for each subject, and calculate the average across subjects per each word. To evaluate the predicted costs, we calculate the Pearson product-moment correlation coefficient ρ between the costs and the average reaction times, with $\rho \in [-1, +1]$ and $\rho = 0$ for uncorrelated variables. This is equilavent to calculating linear regression, as ρ^2 corresponds to the coefficient of determination, i.e., the fraction of variance of the predicted variable explained by the predictor.

2.2 Statistics and Computational Models

Several statistics are calculated for each stimulus word: length, surface frequency, base frequency, morphological family size, and bigram frequency. As logarithmic frequencies often correlate with reaction times better than direct frequencies, we also test those. The computational models examined here give a probability distribution $p(W)$ over the words. Thus, we can use the cost or self-information $-\log p(W)$ to explain the reaction times in a similar manner as with the word frequencies: a high probability is assumed to correlate with a low reaction time.

N-gram Models. We use n-gram models to get a good estimate on how common the form of the word (sequence of letters l_i) is among all the words in the language. An n-gram model of order n is a $(n-1)$:th order Markov model, thus approximating $p(W = l_1 l_2 \ldots l_N)$ as $\prod_{i=1}^{N} p(l_i \,|\, l_{i-n+1} \ldots l_{i-1})$. For estimating the n-gram probabilities $p(l_i \,|\, l_{i-n+1} \ldots l_{i-1})$, the standard techniques include smoothing of the maximum likelihood distributions and interpolation between different lengths of n-grams. We apply one of the state-of-the-art methods, Kneser-Ney interpolation [4], implemented in VariKN toolkit [14].

Morfessor Baseline. Morfessor [6] is a method for unsupervised learning of concatenative morphology. It does not limit the number of morphemes per word, and is thus suitable for modeling complex morphology such as that in Finnish. The basic idea can be explained using the Minimum Description Length (MDL) principle [13], where modeling is viewed as a problem of encoding a data set efficiently in order to transmit it. In two-part MDL coding, one first transmits the model $\mathcal{M}$, and then the data set by referring to the model. Thus the task is to find the model that minimizes the sum of the coding lengths $L(\mathcal{M})$ and $L(\text{corpus}|\mathcal{M})$. In the case of segmenting words into morphs, the model simply consists of a lexicon of unique morphs, and a pointer assigned for each. The corpus is then transmitted by sending the pointer of each morph as they occur in the text. Using $L(X) = -\log p(X)$, the task is equivalent to probabilistic *maximum a posteriori* (MAP) estimation, where $p(\mathcal{M}|\text{corpus})$ is maximized.

In Morfessor Baseline, the lexicon consists of the strings and frequencies of the morphs. The cost of the lexicon increases by the number and length of the morphs. Each pointer in the corpus corresponds to a maximum likelihood probability set according to the morph frequency. Thus, for a known segmentation, the likelihood for corpus is simply the product of the morph probabilities. During training, Morfessor applies a greedy algorithm for finding simultaneously the morph lexicon and a segmentation for the training corpus. After training, a Viterbi-like algorithm can be applied to find the segmentation with the highest probability—the product of the respective morph probabilities—for any single word. For details, see, e.g., [6] and [5].

Morfessor Categories-MAP. The assumption of the independence between the morphs in a word is an obvious problem in Morfessor Baseline. For example, the model gives an equal probability to "s + walk" and "walk + s". The later versions of Morfessor extend the model by adding another layer of representation, namely a Hidden Markov Model (HMM) model of the segments [6]. In Morfessor Categories-MAP, the HMM has four categories (states): prefix, stem, suffix, and non-morpheme. While the model allows hierarchical segmentation to non-morphemes, the final analysis of a word is restricted by the regular expression `(prefix* stem+ suffix*)+`. Context-sensitivity of the model has lead to improved segmentation results when compared to a linguistic gold standard segmentation of words into morphemes [6].

2.3 Data for Learning Computational Models

The main corpus in our experiments is the one used in the Morpho Challenge 2007 competition [8]. It is part of the Wortschatz collection [12] and contains three million sentences collected from World Wide Web. To observe the effect of the training corpus, we also use 30 000, 100 000, 300 000 and one million sentence random subsets of the corpus. In addition, we use three smaller corpora: "Book" (4.4 million words) and "Periodical" (2.1 million words) parts of Finnish Parole corpus [18], subtitles of movies from OpenSubs corpus [19] (3.0 million words), and their combination.

It is often unclear whether intra-word models should be trained on a corpus (word tokens), a word lexicon (types), or something in between. For example, Morfessor Baseline gives segments that correspond better to linguistic morphemes when trained on types rather than tokens [6,5]: with token counts, many inflected high-frequency words are not segmented. Morfessor Categories-MAP, however, is by default trained on tokens [6]: the context-sensitivity of the Markov model reduces the effect of direct corpus frequencies. We compare models trained on types, tokens, and an intermediate approach, where the corpus frequencies c are reduced using a logarithmic function $f(c) = \log(1 + c)$.

3 Results

Table 1 shows the correlations of the different statistics and logarithmic probabilities of the models to the average reaction times for the stimulus words. All values, except for the bigram frequency, showed statistically significant correlation ($p(\rho = 0) < 0.01$). Among the statistics, logarithmic frequencies gave higher

Table 1. Correlation coefficients ρ of different word statistics and models to average human reaction times. Surface frequency I and other statistics are from the Turun Sanomat newspaper corpus. Surface frequency II is from the Morpho Challenge corpus used for training the models. The last row shows correlations for reaction times of individual subjects. The highest correlations are marked with an asterisk.

Word statistics		*Logarithmic*	*Linear*
Surface frequency I		−0.5108	−0.2806
Surface frequency II		−0.5353*	−0.2376
Base frequency		−0.4453	−0.1901
Morphological family size		−0.4233	−0.2916
Bigram frequency		−0.0211	+0.0221
Length (letters)		+0.2180	+0.2158
Length (morphemes)		+0.5417*	+0.5417*
Models	*Types*	*Log-frequencies*	*Tokens*
Letter 1-gram model	+0.1818	+0.1816	+0.1799
Letter 5-gram model	+0.5394	+0.5380	+0.5160
Letter 9-gram model	+0.6952*	+0.6920	+0.6358
Morfessor Baseline	+0.6605	+0.6765*	+0.5817
Morfessor Categories-MAP	+0.6620	+0.6950*	+0.5474
Other	*Minimum*	*Median*	*Maximum*
Reaction times of a single subject	+0.2030	+0.4774	+0.5681*

correlations than linear frequencies, and the highest ones were obtained for the number of morphemes in the word and the surface frequency. Among the models, the n-grams were best trained with word types, while training with the logaritmic frequencies gave the highest correlations for Morfessor. The highest correlation was obtained for the letter 9-gram model trained with word types—any longer n-grams did not improve the results. Categories-MAP correlated almost as well as the 9-gram model, while Baseline did somewhat worse. All of them had markedly higher correlations than the maximum correlation obtained for an single test subject to the average reaction times of the others.

With logarithmic counts, the Categories-MAP model segmented 135 of the 160 inflected nouns, but also 33 of the 160 monomorphemic nouns. The Baseline model segmented less: 39 of the inflected and 5 of the monomorphemic nouns.

Figure 1 shows how the reaction times and probabilities given by Categories-MAP model match for individual stimulus words. Observing the words that have poor match between the predicted difficulty and reaction time led us to suspect that some of the unexplained variance is due to a training corpus that does not match the material that humans are exposed to. Thus we next studied the effect of the training corpus for the morphological models (Fig. 2). Increasing the amount of word types in the corpus clearly improved the correlation between model predictions and measured reaction times. However, the data from books, periodicals and subtitles gave usually higher correlations than the same amount of the Morpho Challenge data.

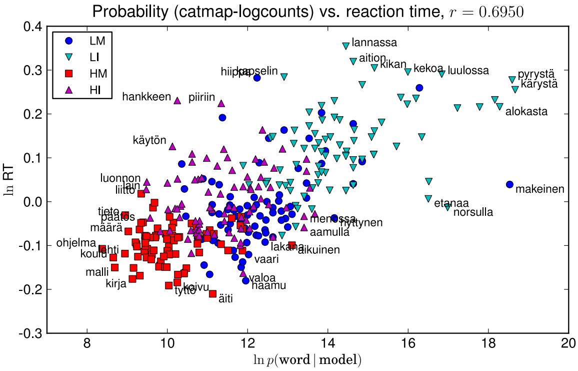

Fig. 1. Scatter plot of reaction times and log-probabilities from Morfessor Categories-MAP. The words are divided into four groups: low-frequency monomorphemic (LM), low-frequency inflected (LI), high-frequency monomorphemic (HM), and high-frequency inflected (HI). Words that have faster reaction times than predicted are often very concrete and related to family, nature, or stories: *tyttö* (girl), *äiti* (mother), *haamu* (ghost), *etanaa* (snail + partitive case), *norsulla* (elephant + adessive case). Words that have slower reaction times than predicted are often more abstract or professional: *ohjelma* (program), *tieto* (knowledge), *hankkeen* (project + genitive case), *käytön* (usage + genitive case), *hiippa* (miter), *kapselin* (capsule + genitive case).

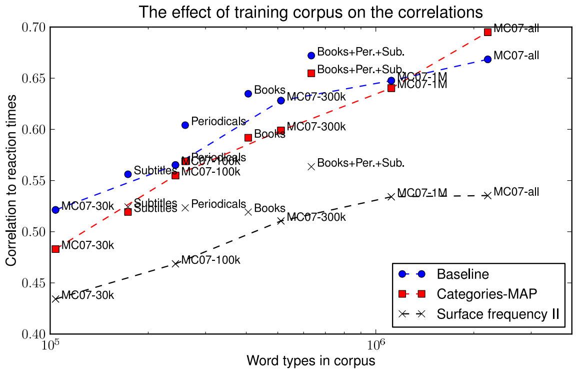

Fig. 2. The effect of training corpus on correlations of Morfessor Baseline (blue circles), Categories-MAP (red squares), and logarithmic surface frequencies (black crosses). The dotted lines show the results on subsets of the same corpus. Unconnected points show the results using different types of corpora.

4 Discussion

We studied how language models trained on unannotated textual data can predict human reaction times for inflected and monomorphemic Finnish words in a lexical decision task. Three models, the letter-based 9-gram model and the

Morfessor Baseline and Categories-MAP models, provided not only higher correlations than the simple statistics of words previously identified as important factors affecting the recognition times in morphologically complex words (cf. [16,1,2]), but also higher than the correlations of reaction times of individual subjects to the average times of the others. The level of correlation was surprisingly high especially because the training corpus is likely to differ from the material humans encounter during their course of life. Based on the results using several training corpora, we assume that even higher correlations would be obtained with more realistic training data.

The highest correlations were obtained for the letter 9-gram model. However, its number of parameteres—almost 6 million n-gram probabilities—was very large. As the estimates of the word probabilities are very precise, we assume that they are good predictors especially for early visual processing stages.

The Categories-MAP model had almost as high correlation as the 9-gram model with much fewer parameters (178 000 transition and emission probabilities). It has three important aspects: First, it applies morpheme-like units instead of words or letters. Second, it finds units that provide a compact representation for the data. Third, the model is context-sensitive: the cost of next unit depends on the previous unit. It is still unclear which contributes more to the high correlations: the morpheme lexicon learned by minimizing the description length, or the underlying probabilistic model. One way to study this question further is to apply a similar model to a linguistic morphological analysis of a corpus.

While behavioral reaction times necessarily incorporate multiple processing stages, brain activation measures could provide markedly more precise markers of the different stages of visual word processing. At the level of the brain, effects of morphology have been previously detected in neural responses that have been associated with later stages of word recognition such as lexical-semantic, phonological and syntactic processing [9,20]. Future work includes finding out whether the predictive power of the models stems from some of these stages, or from an earlier one related to the processing of visual word forms.

Acknowledgments. This work was funded by Academy of Finland, Graduate School of Language Technology in Finland, Sigrid Jusélius Foundation, Finnish Cultural Foundation, and Stiftelsen för Åbo Akademi.

References

1. Alegre, M., Gordon, P.: Frequency effects and the representational status of regular inflections. Journal of Memory and Language 40, 41–61 (1999)
2. Bertram, R., Baayen, R., Schreuder, R.: Effects of family size for complex words. Journal of Memory and Language 42, 390–405 (2000)
3. Butterworth, B.: Lexical representation. In: Butterworth, B. (ed.) Language Production, pp. 257–294. Academic Press, London (1983)
4. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech & Language 13(4), 359–393 (1999)

5. Creutz, M., Lagus, K.: Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Tech. Rep. A81. Publications in Computer and Information Science. Helsinki University of Technology (2005)
6. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing 4(1) (January 2007)
7. Karlsson, F.: Suomen kielen äänne- ja muotorakenne (The Phonological and Morphological Structure of Finnish). Werner Söderström, Juva (1983)
8. Kurimo, M., Creutz, M., Varjokallio, M.: Morpho challenge evaluation using a linguistic gold standard. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 864–872. Springer, Heidelberg (2008)
9. Lehtonen, M., Cunillera, T., Rodríguez-Fornells, A., Hultén, A., Tuomainen, J., Laine, M.: Recognition of morphologically complex words in Finnish: evidence from event-related potentials. Brain Research 1148, 123–137 (2007)
10. Lehtonen, M., Laine, M.: How word frequency affects morphological processing in monolinguals and bilinguals. Bilingualism: Language and Cognition 6, 213–225 (2003)
11. Niemi, J., Laine, M., Tuominen, J.: Cognitive morphology in Finnish: foundations of a new model. Language and Cognitive Processes 9, 423–446 (1994)
12. Quasthoff, U., Richter, M., Biemann, C.: Corpus portal for search in monolingual corpora. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, pp. 1799–1802 (2006)
13. Rissanen, J.: Modeling by shortest data description. Automatica 14, 465–471 (1978)
14. Siivola, V., Hirsimäki, T., Virpioja, S.: On growing and pruning Kneser-Ney smoothed n-gram models. IEEE Transactions on Audio, Speech & Language Processing 15(5), 1617–1624 (2007)
15. Soveri, A., Lehtonen, M., Laine, M.: Word frequency and morphological processing revisited. The Mental Lexicon 2, 359–385 (2007)
16. Taft, M.: Recognition of affixed words and the word frequency effect. Memory and Cognition 7, 263–272 (1979)
17. Taft, M.: Morphological decomposition and the reverse base frequency effect. The Quarterly Journal of Experimental Psychology A 57, 745–765 (2004)
18. The Department of General Linguistics, University of Helsinki and Research Institute for the Languages of Finland (gatherers): Finnish Parole Corpus (1996–1998), available through CSC, `http://www.csc.fi/`
19. Tiedemann, J.: News from OPUS — A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing, vol. 5, pp. 237–248. John Benjamins, Amsterdam (2009)
20. Vartiainen, J., Aggujaro, S., Lehtonen, M., Hultén, A., Laine, M., Salmelin, R.: Neural dynamics of reading morphologically complex words. NeuroImage 47, 2064–2072 (2007)