# MORPHOLOGICALLY MOTIVATED LANGUAGE MODELS IN SPEECH RECOGNITION

*Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo*

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FI-02015 HUT, Finland, teemu.hirsimaki@hut.fi

## ABSTRACT

Language modelling in large vocabulary speech recognition has traditionally been based on words. A lexicon of the most common words of the language in question is created and the recogniser is limited to consider only the words in the lexicon. In Finnish, however, it is more difficult to create an extensive lexicon, since the compounding of words, numerous inflections and suffixes increase the number of commonly used word forms considerably. The problem is that reasonably sized lexica lack many common words, and for very large lexica, it is hard to estimate a reliable language model.

We have previously reported a new approach for improving the recognition of inflecting or compounding languages in large vocabulary continuous speech recognition tasks. Significant reductions in error rates have been obtained by replacing a traditional word lexicon with a lexicon based on morpheme-like word fragments learnt directly from data. In this paper, we evaluate these so called statistical morphs further, and compare them to grammatical morphs and very large word lexica using n-gram language models of different orders. When compared to the best word model, the morph models seem to be clearly more effective with respect to entropy, and give 30% relative error-rate reductions in a Finnish recognition task. Furthermore, the statistical morphs seem to be slightly better than the rule-based grammatical morphs.

## 1. INTRODUCTION

Automatic speech recognition is based on acoustics, but modern speech recognition systems rely heavily on models of the language too. In practice, all speech recognition systems do some kind of search, in which different sentences are hypothesised and their probability is computed using the acoustic models and language models. In the end, the hypothesis giving the highest probability is chosen as the recognition output. Because all possible sentences of any language obviously can not be tried and evaluated, the most improbable hypotheses must be pruned away at an early stage, and the computation is concentrated on the most probable hypotheses.

Especially in the recognition of English speech, a traditional way to limit the search space is to construct a lexicon of the most common words, and let the recogniser consider words from the lexicon only. Typically the size of the lexicon is something between 10 000 and 60 000 words. Restricting the recogniser to certain words naturally poses the problem that the words outside the lexicon can not be recognised correctly. These words are called out-of-vocabulary (OOV) words in speech recognition literature.

In English, the problem of OOV words is not so severe, but in Finnish, it is not reasonable to build an extensive lexicon for general speech. Because compound words and inflections are very common in Finnish, and words are often formed by adding a few suffixes to a base form, the number of distinct word forms is very large. Using larger and larger lexica makes the OOV words less common, but at the same time, it also complicates the use of language models. The same OOV problem can also be seen in other highly inflecting languages like Turkish and Hungarian, and compounding languages like German, Greek and Swedish, for example.

Several approaches to tackle the problem have been proposed in the literature. First, there are approaches that try to expand the vocabulary with the most frequent word forms either dynamically or statically, e.g., German [1, 2] and Finnish [3]. A different promising direction is to abandon the word as lexical unit and split words into smaller word fragments. Then a large number of words can be created with a reasonably sized fragment lexicon. The proposed methods range from hand-crafted rules to unsupervised data-driven methods for different languages, e.g., German and Finnish [4], Korean [5], Greek [6], Hungarian [7], and Dutch [8].

We have earlier used an unsupervised data-driven algorithm [9] to find an efficient set of word fragments for speech recognition. The fragments produced by our algorithm resemble grammatical morphemes, which are the smallest meaning-bearing units in language, and we call them *statistical morphs*. In comparison with words and syllables, the morphs have given clear error rate reductions in a Finnish unlimited vocabulary continuous recognition task [10]. The method is language independent, and has also given good results for Turkish [11].

In this paper, we develop and evaluate the statistical morphs further. The important questions addressed in the experiments are the following: Are the error rate reductions obtained with statistical morphs only due to the fact, that the OOV problem is avoided, because any word form can be formed from smaller units? Or would other ways to split words into fragments give good results too? To study the issue, we have also built other language models that use different set of words and word fragments, and can form any word form from the fragments. The models we compare to the statistical morphs are based on two lexica: huge word lexica extended with Finnish phonemes, and morphs based on a grammatical analysis, also extended with phonemes. The performance of the models are evaluated in cross-entropy and speech recognition experiments.

## 2. LEXICA AND LANGUAGE MODELS

We investigate three different types of lexical units: (i) *statistical morphs* that have been found efficient in Finnish speech recognition; (ii) *words* extended with phonemes as sub-word units; (iii) *grammatical morphs* that illustrate how a linguistic hand-made model can be applied to produce word fragments.

Because the optimal size of the lexicon may vary for different lexical units, we have generated lexica of different sizes. On the one hand, we have aimed at lexica containing the same number of units regardless of the type of unit. This has resulted in a word lexicon containing approximately 69 000 words, a grammatical morph lexicon containing about 79 000 grammatical morphs, and a statistical morph lexicon containing 66 000 morphs. On the other hand, we have aimed at optimal performance for the approaches, which has resulted in a word lexicon of 410 000 words and a statistical morph lexicon of 26 000 morphs. The number of grammatical morphs was fixed, since these morphs were produced using a rule set.

### 2.1. Statistical morphs

The statistical morphs are found using the *Recursive MDL* algorithm [9], which learns a model inspired by the Minimum Description Length (MDL) principle. A more detailed description of the algorithm is presented in a technical report [12], and the implementation is publicly available[1]. The basic idea is to run the algorithm on a large text corpus, and the algorithm tries to find a morph lexicon that encodes the corpus efficiently, but is still compact itself. In practice, this principle splits words in fragments if the fragments are useful in building other common words. The rarest words end up being split in many fragments, while very common words remain unsplit.

Unlike the original version of the algorithm [9], we do not use the corpus as such as training data for the algorithm, but a word list containing one occurrence of each word in the corpus. In the original approach, large training corpora lead to large morph lexica, since the algorithm needs to find a balance between the two in its attempt to obtain the globally most concise model. By choosing only one occurrence of every word form as training data, the optimal balance occurs at a smaller morph lexicon, while still preserving the ability to recognise good morphs, which are common strings that occur in different combinations with other morphs. A morph lexicon containing 66 000 morphs was produced in this way. Another even smaller morph lexicon (26 000 morphs) was obtained by training the algorithm on a word list where word forms occurring less than three times in the corpus were filtered out. This approach is motivated by the fact that many word forms, that occur only a few times in the corpus, might be noise (such as misspellings and foreign words) and their removal might increase the robustness of the algorithm.

Once the lexicon is ready, every word form in the corpus is segmented into the most likely morph sequence using Viterbi search. Finally, n-gram language models are estimated over the segmented corpus. As words can consist of multiple morphs, word boundaries need to be modelled explicitly. The lexicon contains a special word boundary morph, which terminates each word.

### 2.2. Words

As mentioned in the introduction, OOV words become a problem when the lexicon is constructed of unsplit word forms. To see if this problem could be alleviated in a simple way, we have tried adding phonemes to the lexicon. As usual, the most common words are selected into the lexicon directly, but instead of discarding the remaining OOV words, they are split into phonemes so that it is possible to construct any word form by concatenating phonemes. N-gram language models are estimated as usual over the training corpus, where the rare word forms have been split into phonemes. For our larger word lexicon of 410 000 words, this means that 5% of the words in the training corpus are split into phonemes. In the data used for testing the speech recogniser, nearly 8% of the words are split.

As this combination of words and phonemes avoids OOV words, it can be compared fairly to the statistical morphs. Note, that the Finnish orthography and pronunciation have a close correspondence, which makes it rather straightforward for a recognition application to rejoin and correctly spell out words that have been built by concatenating phonemes.

Unlike in the statistical morph model, word breaks are modelled so that we have two variants of each phoneme in the lexicon, one for occurrences at the end of a word, and one for other cases. Each unsplit word is assumed implicitly to end in a word break.

### 2.3. Grammatical morphs

In order to obtain a segmentation of words into grammatical morphs, each word form was run through a morphological analyser[2] based on the two-level morphology of Koskenniemi [13]. The output of the analyser consists of the base form of the word together with grammatical tags indicating, e.g., part-of-speech, number and case. Boundaries between the constituents of compound words are also marked. We have created a rule set that

[1]http://www.cis.hut.fi/projects/morpho/

[2]Licensed from Lingsoft, Inc.: http://www.lingsoft.fi/

| | |
|---|---|
| Statist. morphs (26k) | tuore mehu asema # al oitti # omena mehu n # purista misen # pyy nik illä # |
| Words (410k) | t u o r e m e h u a s e m a# aloitti# omenamehun# puristamisen# pyynikillä# |
| Grammatical morphs | tuore mehu asema # aloitt i # omena mehu n # purista mise n # p yy n i k i ll ä # |
| Literal translation | fresh juice station # start -ed # apple juice of # press -ing # Pyynikki in # |

Table 1. A phrase of the training corpus segmented using different lexical units. (An English translation reads: *"A juice factory [has] started to press apple juice in Pyynikki".*) The lexical units are separated by space. Word breaks are indicated by a number sign (#). In case of the word model, the word breaks are part of other lexical units, otherwise they are units of their own.

| | Gramm. (79k) | Stat. (26k) | Stat. (66k) | Word (69k) | Word (410k) |
|---|---|---|---|---|---|
| Gramm. (79k) | 100% | 41% | 37% | 20% | 19% |
| Stat. (26k) | | 100% | 72% | 23% | 23% |
| Stat. (66k) | | | 100% | 34% | 35% |
| Word (69k) | | | | 100% | 85% |
| Word (410k) | | | | | 100% |

Table 2. Pairwise similarity of the segmentation of the test set obtained with different models. Each figure is the percentage of the test set that is segmented into identical morphs when using two different models, i.e., the percentage of phonemes that are covered with identical morphs.

processes the output of the analyser and produces a grammatical morph segmentation of the words in the corpus. The rules in our rule set are close to the morphological description for Finnish given in [14].

A slightly newer version of the grammatical morph segmentation, called *Hutmegs* (Helsinki University of Technology Morphological Evaluation Gold Standard), is publicly available for research purposes [15]. For full functionality, an inexpensive license must additionally be purchased from Lingsoft, Inc.

Words not recognised by the morphological analyser are treated as OOV words in the word model and split into individual phonemes. Such words make up 4% of all the words in the training corpus, but only 0.3% of the words in the test data. N-gram language models are estimated over the training corpus, and just like in statistical morph model, word boundaries are modelled explicitly as separate units.

### 2.4. Comparison of the segmentations

Figure 1 shows the splittings of the same Finnish example sentence using the three different lexicon types. The Finnish word for "juice factory" is rare and therefore it is split into phonemes in the word model, whereas the place name "Pyynikki" is unknown to the morphological analyser. The statistical morph model needs not resort to individual phonemes very often, even when representing rare words, for instance proper names.

Even if the lexicons are different, there is some overlap between their morph inventories. To measure the overlap, we segmented the test data using the five lexicons and studied how often two different models segmented the data in identical morphs. Table 2 shows, for each pair of models, the percentage of phonemes covered by identical morphs. It can be seen that the word lexicons produce very similar segmentations (85% overlap) compared to each other, although the size of the lexicons is very different. The same applies to the two statistical morph lexicons (72% overlap). This suggests that the lexicons differ mostly in how they model rare events. Compared to the segmentation obtained with the grammatical morphs, the statistical morphs produce more similar segmentations (around 40%) than word lexicons do (around 20%).

## 3. EXPERIMENTS

### 3.1. Data

In the experiments, we used the same data as in our previous work [10]. The lexical units and language models were trained from a corpus of 36 million words from the Finnish News Agency (newswires) and the Finnish IT center (books, newspapers, magazines).

The speech data was a talking book read by a female speaker. 12 hours of the book were used for training the acoustic models, 21 minutes for tuning decoder parameters and 26 minutes for testing. The transcription of the first 12 hours of the book was used as the test set for the language model entropy tests.

### 3.2. Language models and cross-entropy

For each lexicon type, we trained n-gram language models of order 2–7. The SRI-toolkit [16] was used with Kneser-Ney smoothing. Numbers and abbreviations were automatically expanded to words and foreign names were converted to their phonetic representations.[3] These forms were used in the evaluation of both the cross-entropy and speech recognition results.

In order to measure the quality of language models before running speech recognition tests, it is common to measure the modelling performance of the models on text

[3]We are grateful to Mr. Sami Virpioja for giving technical help with the SRI-toolkit, and Mr. Nicholas Volk for kindly providing the transcription software: http://www.ling.helsinki.fi/suopuhe/lavennin/

data. The most common measures are *cross-entropy* and *perplexity* that are based on the probability of a test corpus, that has not been used in training the models. The cross-entropy $H_M(T)$ of the model $M$ on the data $T$ is given by

$$H_M(T) = -\frac{1}{W(T)} \log_2 P(T|M) \qquad (1)$$

where $W(T)$ is number of words in the test data. The cross-entropy tells the minimum number of bits needed to encode each word on average [17]. Usually, the data probability $P(T|M)$ is decomposed into probabilities of words, but we decompose it into probabilities of word fragments or morphs:

$$P(T|M) = \prod_{i=1}^{F_M(T)} P(f_i|f_{i-1}, \ldots, f_1; M) \qquad (2)$$

where $F_M(T)$ is the number of word fragments and $f_i$ are the fragments according to model $M$. And as usual when n-gram models are in question, only a few preceding words are taken into account instead of whole history $(f_{i-1}, \ldots, f_1)$. Note that the metric is normalised by the number of words in the test data. Thus, it is fair even if the models use different fragments to compute word probabilities.

The other common measure, *perplexity*, is very closely related to cross-entropy, and it is defined as follows:

$$\mathrm{Perp}_M(T) = \left( \prod_{i=1}^{W_T} P(w_i|w_{i-1}, \ldots, w_1; M) \right)^{-\frac{1}{W_T}}. \qquad (3)$$

From the above, it is easy to see that the relation to cross-entropy is given by

$$\mathrm{Perp}_M(T) = P(T|M)^{-\frac{1}{W_T}} \qquad (4)$$
$$= 2^{H_M(T)} \qquad (5)$$

We have measured cross-entropy in the experiments. Figure 1 on the next page shows the cross-entropies of our models with respect to the model sizes. It can be seen that for smaller models, the morpheme-based language models offer a significantly more effective way of modelling the language. In addition to the reported language model sizes, large lexica consume more memory in the decoding process.

### 3.3. Speech recognition experiments

The cross-entropy experiments only measure the general modelling power of the language models, and do not predict very accurately how well the models will perform in speech recognition tasks. This is especially the case when the language models in question are estimated over different sets of sub-word units. Thus, it is important to evaluate the models in real speech recognition experiments too.

Next the speech recognition system used in the experiments is described briefly. A more detailed description of the system can be found in [18].

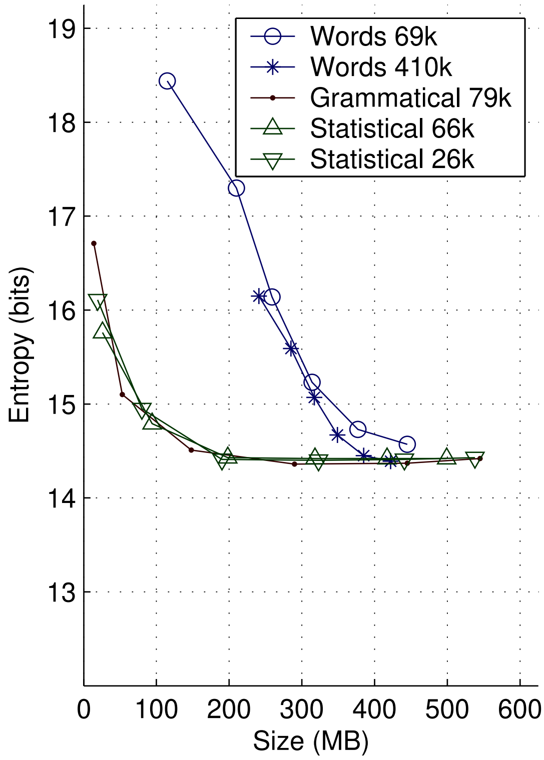


Figure 1. The cross-entropies and model sizes for different lexicon types. N-gram models of order 2–7 were tested.

The acoustic phoneme models of our recogniser were Hidden Markov Models with Gaussian mixture emission probabilities. Compared to our previous experiments [10], two improvements were made to the acoustic models: A global linear transform optimised in maximum likelihood sense was used to make the feature components maximally uncorrelated for each diagonal Gaussian mixture component. In addition, phoneme durations were modelled. During recognition, the acoustic probability of each hypothesis was updated according to how the recognised phone durations fit the trained duration distributions. For duration modelling, gamma distributions were used [19]. The duration modelling is important for Finnish, since each phoneme has a long and a short variant. In this experiment, we used monophones instead of triphones. Since our decoder does not handle phoneme contexts across lexical units, this was the fairest way to compare the language models based on different lexical units.

Our one-pass decoder uses a stack decoding approach by storing hypotheses in frame-wise stacks. The idea is to make a local acoustic search separately for hypotheses ending at different time frames. The language model probabilities are added when the hypotheses are inserted in the stacks. The approach makes it possible to use different language models easily without affecting the acoustic search.

The phoneme error rates (PHER) of the recognition experiments are shown in Figure 2. For each lexicon type,

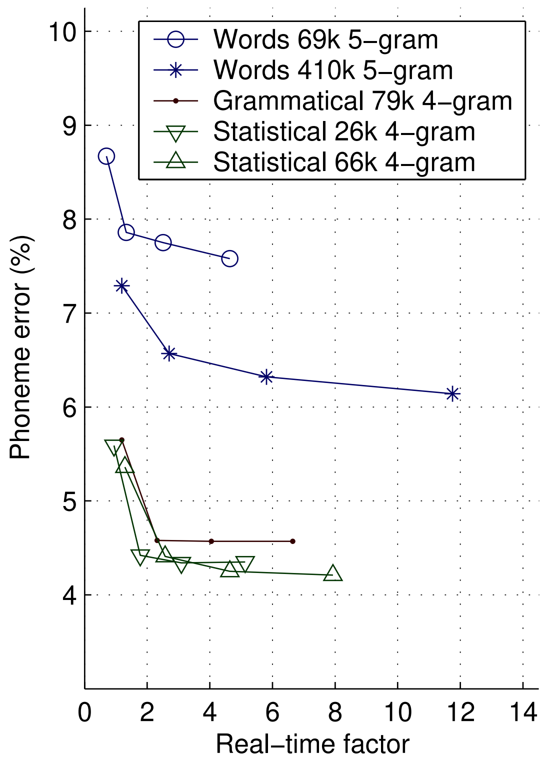


Figure 2. Recognition results. For each lexicon type, the phoneme error curve of the best n-gram model is shown (orders 3–5 were tested). For each model, four different decoder pruning settings were used, giving varying real-time factors.

n-gram language models of order 3–5 were used, and four different decoder pruning settings were tested in order to study the behaviour at different decoding speeds. The figure shows only the curve of the best language model order for each lexicon type. The word error rates (WER) behave similarly. For the best morph model, the PHER 4.2% corresponds to WER 21%, and for the best word model the PHER 6.1% corresponds to WER 30%.

## 4. DISCUSSION

In the cross-entropy tests (Fig. 1), the word models reach the performance of the morph models, when the order of the n-gram models is increased. However, the same behaviour is not observed in the recognition results (Fig. 2). One might argue that this is partly due to the decoder approach. Since the language model probabilities are taken into account only at the ends of the lexical units, the longer word models are pruned more easily. But relaxing the pruning settings of the decoder does not seem to help the word model, so other explanations for the difference must be sought.

One reason is probably the number of words that the models based on words and grammatical morphs have to split into phonemes. As reported in Sections 2.2 and 2.3, the proportion of OOV words in the training corpus is roughly the same for both the large word model and the grammatical morph model. But whereas the OOV rate of the test data is only 0.3% for the grammatical morphs, it is as much as 8% for the words. Even if rare word forms can be built from phonemes (giving a fair entropy comparison), this does not help the word model considerably in the actual recognition task.

As far as the statistical morphs are concerned, it is interesting that the actual number of morphs in the lexicon does not seem to affect the results very much. What seems to be important is that words are split into more common parts for which more occurrences and thus better probability estimates can be obtained. At the same time, over-fragmentation into individual phonemes is not as common as in the other models. Over-fragmentation apparently causes problems that cannot be remedied using Kneser-Ney smoothing, even though this type of smoothing is known to perform better than other well-known smoothing techniques in language modelling [20].

It would be interesting to study further, how small a morph lexicon can be used before the performance starts to degrade. It is likely, however, that the optimal units for language modelling do have a connection to morphemes or morpheme-like units, which function as rather independent entities in the syntax, and also the semantics, of a language. As a basis for the representation of linguistic knowledge, such units seem well motivated, and might also be very useful in language models that try to capture the semantic and syntactic dependencies of the language better than the n-gram model, such as structured language models [21].

## 5. CONCLUSION

To sum up, finding a balanced set of lexical units is important in speech recognition of inflecting and compounding languages. Both the grammatical and statistical morpheme-like word fragments seem to be good choices for representing a very large vocabulary efficiently with a reasonable number of lexical units. The statistical morphs have the additional advantage of being produced in an unsupervised and language independent manner.

## 6. ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland in the projects *New information processing principles* and *New adaptive and learning methods in speech recognition*. Funding was also provided by the Finnish National Technology Agency (TEKES) and the Graduate School of Language Technology in Finland. The audio data was provided by the Finnish Federation of the Visually Impaired, and Departments of Speech Science and General Linguistics of the University of Helsinki. We are also grateful to the Finnish news agency (STT) and the Finnish IT center for science (CSC) for the text data. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. We acknowledge that access rights to data

and other materials are restricted due to other commitments.

## 7. REFERENCES


[1] P. Geutner, M. Finke, and P. Scheytt, "Adaptive vocabularies for transcribing multilingual broadcast news," in *Proc. ICASSP*, 1998, pp. 925–928.

[2] Kevin McTait and Martine Adda-Decker, "The 300k LIMSI German broadcast news transcription system.," in *Proc. Eurospeech*, 2003, pp. 213–216.

[3] Vesa Siivola, Mikko Kurimo, and Krista Lagus, "Large vocabulary statistical language modeling for continuous speech recognition in Finnish," in *Proc. Eurospeech*, 2001, pp. 737–740.

[4] Jan Kneissler and Dietrich Klakow, "Speech recognition for huge vocabularies by using optimized subword units," in *Proc. Eurospeech*, 2001, pp. 69–72.

[5] Young-Hee Park, Dong-Hoon Ahn, and Minhwa Chung, "Morpheme-based lexical modeling for Korean broadcast news transcription," in *Proc. Eurospeech*, 2003, pp. 1129–1132.

[6] Dimitros Oikonomidis and Vassilios Digalakis, "Stem-based maximum entropy language models for inflectional languages," in *Proc. Eurospeech*, 2003, pp. 2285–2288.

[7] Máté Szarvas and Sadaoki Furui, "Evaluation of the stochastic morphosyntactic language model on a one million word Hungarian task," in *Proc. Eurospeech*, 2003, pp. 2297–2300.

[8] Roeland Ordelman, Arjan van Hessen, and Franciska de Jong, "Compound decomposition in Dutch large vocabulary speech recognition," in *Proc. Eurospeech*, 2003, pp. 225–228.

[9] Mathias Creutz and Krista Lagus, "Unsupervised discovery of morphemes," in *Proc. of the Workshop on Morphological and Phonological Learning of ACL-02*, 2002, pp. 21–30.

[10] Vesa Siivola, Teemu Hirsimäki, Mathias Creutz, and Mikko Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proc. Eurospeech*, 2003, pp. 2293–2296.

[11] Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo, and Mathias Creutz, "On lexicon creation for Turkish LVCSR," in *Proc. Eurospeech*, 2003, pp. 1165–1168.

[12] Mathias Creutz and Krista Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Publications in Computer and Information Science A81, Helsinki University of Technology, Mar. 2005.

[13] K. Koskenniemi, *Two-level morphology: A general computational model for word-form recognition and production*, Ph.D. thesis, University of Helsinki, 1983.

[14] Lauri Hakulinen, *Suomen kielen rakenne ja kehitys (The structure and development of the Finnish language)*, Kustannus-Oy Otava, 4 edition, 1979.

[15] Mathias Creutz and Krister Lindén, "Morpheme segmentation gold standards for Finnish and English," Tech. Rep. A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004, URL: http://www.cis.hut.fi/projects/morpho/.

[16] Andreas Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[17] Stanley F. Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, no. 4, pp. 359–393, 1999.

[18] Teemu Hirsimäki and Mikko Kurimo, "Decoder issues in unlimited Finnish speech recognition," in *Proc. of the 6th Nordic Signal Processing Symposium (Norsig)*, 2004, pp. 320–323.

[19] Janne Pylkkönen and Mikko Kurimo, "Using phone durations in Finnish large vocabulary continuous speech recognition," in *Proc. of the 6th Nordic Signal Processing Symposium (Norsig)*, 2004, pp. 324–327.

[20] Joshua T. Goodman, "A bit of progress in language modeling," *Computer Speech and Language*, vol. 15, pp. 403–434, 2001.

[21] Ciprian Chelba and Frederick Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, no. 4, pp. 283–332, 2000.