Morph-Based Speech Recognition and Modeling of Out-of-Vocabulary Words Across Languages

MATHIAS CREUTZ, TEEMU HIRSIMÄKI, MIKKO KURIMO, ANTTI PUURULA, JANNE PYLKKÖNEN, VESA SIIVOLA, MATTI VARJOKALLIO
Helsinki University of Technology
EBRU ARISOY, MURAT SARAÇLAR
Boğaziçi University, Istanbul
and
ANDREAS STOLCKE
SRI International, Menlo Park
International Computer Science Institute, Berkeley

We explore the use of morph-based language models in large-vocabulary continuous speech recognition systems across four so-called "morphologically rich" languages: Finnish, Estonian, Turkish, and Egyptian Colloquial Arabic. The morphs are subword units discovered in an unsupervised, data-driven way using the *Morfessor* algorithm. By estimating n-gram language models over sequences of morphs instead of words, the quality of the language model is improved through better vocabulary coverage and reduced data sparsity. Standard word models suffer from high out-of-vocabulary (OOV) rates, whereas the morph models can recognize previously unseen word forms by concatenating morphs. It is shown that the morph models do perform fairly well on OOVs without compromising the recognition accuracy on in-vocabulary words. The Arabic experiment constitutes the only exception, since here the standard word model outperforms the morph model. Differences in the data sets and the amount of data are discussed as a plausible explanation.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—Speech recognition and synthesis, Language models; I.5.1 [**Pattern Recognition**]: Models—statistical

General Terms: Algorithms, Experimentation, Languages, Performance

Additional Key Words and Phrases: Highly inflecting and compounding languages, morphologically rich languages, subword-based language modeling, morpheme, LVCSR, Morfessor, n-gram models, Finnish, Estonian, Turkish, Egyptian Colloquial Arabic

1. INTRODUCTION

As automatic speech recognition systems are being developed for an increasing number of languages, there is growing interest in language modeling approaches that are suitable for so-called "morphologically rich" languages. In these languages, the number of possible word forms is very large because of many productive morphological processes; words are

Address of corresponding author: Mathias Creutz, Helsinki University of Technology, Adaptive Informatics Research Centre, P.O. Box 5400, FIN-02015 TKK, Finland. E-mail: mathias.creutz@tkk.fi

formed through extensive use of, e.g., inflection, derivation and compounding (such as the English words ‘rooms’, ‘roomy’, ‘bedroom’, which all stem from the noun ‘room’).

English is a fairly “morphologically poor” language, and consequently language modeling on the word level has proven successful, or at least satisfactory. The recognition vocabulary consists of a list of word forms observed in the training text, and n-gram language models are estimated over sequences of words. In comparison to other “morphologically richer” languages, for English this approach does not typically lead to huge vocabularies or severe sparsity problems, nor to a high proportion of out-of-vocabulary (OOV) words in a held-out independent test text. A high OOV rate would be problematic, since OOVs are words that are not present in the recognition vocabulary, and thus can never be recognized correctly by the speech recognizer. In addition, the presence of out-of-vocabulary words may cause the misrecognition of several in-vocabulary words adjacent to the OOVs [Klakow et al. 1999; Hacioglu et al. 2003]. Common wisdom, which is supported by experiments [Bisani and Ney 2005], says that each OOV word causes between 1.5 and 2 word errors, on average.

The applicability of the word-based approach to morphologically richer languages has been questioned. In highly compounding languages, such as the Germanic languages German, Dutch and Swedish, decomposition of compound words can be carried out in order to reduce the vocabulary size and the sparsity of the text data; see, e.g., Berton et al. [1996], Larson et al. [2000], Ordelman et al. [2003]. Highly inflecting languages are found, e.g., among the Slavic, Romance, Turkic, and Semitic language families. Language models incorporating morphological knowledge about these languages have been applied, such as two-pass recognition of Serbo-Croatian, where word inflections are included in the second pass [Geutner et al. 1998], Factored Language Models for Arabic [Bilmes and Kirchhoff 2003; Kirchhoff et al. 2006], decomposition into subword units for Turkish [Hacioglu et al. 2003; Arısoy et al. 2006; Kurimo et al. 2006b], and subword modeling combined with two-pass recognition for Turkish [Arısoy and Saraçlar 2006]. A further category comprises languages that are both highly inflecting and highly compounding, such as the Finno-Ugric languages Finnish and Estonian. Here decomposition of words into smaller subword units has been performed, mainly using statistical methods. Language models have then been trained over sequences of such subword units instead of entire words, e.g., Kneissler and Klakow [2001], Hirsimäki et al. [2006], Kurimo et al. [2006].

What all the morphology-modeling approaches have in common is the aim to reduce the OOV rate as well as data sparsity, thereby obtaining more effective language models for morphologically rich languages. Intuitively speaking, language modeling at the subword level makes sense. However, obtaining considerable improvements in speech recognition accuracy seems hard, as is demonstrated by the fairly meager improvements (1–4 % relative) over standard word-based models accomplished by, e.g., Berton et al. [1996], Ordelman et al. [2003], Kirchhoff et al. [2006], Whittaker and Woodland [2000], Kwon and Park [2003], and Shafran and Hall [2006] for Dutch, Arabic, English, Korean, and Czech, or even the worse performance reported by Larson et al. [2000] for German and Byrne et al. [2001] for Czech. Nevertheless, clear improvements over a word baseline have been achieved for Serbo-Croatian [Geutner et al. 1998], Finnish, Estonian [Kurimo et al. 2006] and Turkish [Kurimo et al. 2006b].

This paper analyzes subword language models (LMs) across languages. Speech recognition systems for four languages are studied: Finnish, Estonian, Turkish, and the dialect of Arabic spoken in Egypt, Egyptian Colloquial Arabic (ECA). All four languages are considered "morphologically rich", but the experiments carried out show differences in their behavior: The applied subword LMs are clearly beneficial for Finnish and Estonian, mostly beneficial for Turkish, and slightly detrimental for ECA. The current paper attempts to discover explanations for these differences.

In particular, the focus is on the analysis of OOVs: A perceived strength of subword models, when contrasted with word models, is that subword models can generalize to previously unseen word forms by recognizing them as sequences of shorter familiar word fragments. But does this theory work in practice? Another question concerns the structure of the language studied: Can the OOV and data sparsity problem be solved by other means, through a drastic increase of the amount of LM training data? Moreover, what are the likely effects of spontaneous vs. planned (read) speech?

Speech recognizers are utterly complex systems with a multitude of variables. The conclusions drawn in the current paper rely on experiments performed on different platforms with different acoustic conditions and different amounts of training data. All variables cannot be controlled when making comparisons across languages. The good thing is that this variability may be a strength in that it is possible to demonstrate the platform-independent robustness of the tested methods. Nonetheless, the discussion is bound to bear some speculative traits, which, however, we hope will not repel the reader, and will leave room for future discussion.

1.1 Structure of the paper

The remainder of the paper is structured as follows: Section 2 outlines the algorithm *Morfessor*, which has been used for creating the set of subword units utilized as the lexicon of the speech recognition systems. Section 3 briefly describes the recognition systems and data sets used and provides a thorough analysis of the obtained speech recognition results. In Section 4, the discussion turns to some alternative approaches and additional remarks. Section 5 summarizes the main findings of the paper.

2. MORFESSOR

Morfessor is an unsupervised data-driven method for the segmentation of words into morpheme-like units. The general idea behind the Morfessor model is to discover as compact a description of the input text data as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs* and the words are then represented as a concatenation of morphs, e.g., 'hand, hand+s, left+hand+ed, hand+ful'.

An optimal balance is sought between compactness of the *morph lexicon* versus the compactness of the representation of the *corpus*. The morph lexicon is a list of all distinct morphs (e.g., 'hand, s, left, ed, ful'). The representation of the corpus can be seen as a sequence of pointers to entries in the morph lexicon; e.g. the word 'lefthanded' is represented as three pointers to morphs in the lexicon.

A very compact lexicon could consist of the individual letters of the language. However, this would result in a very expensive representation of the corpus, since every word would be broken down into as many morphs as the number of letters it contains. The opposite situation consists of having a short representation of the corpus (e.g., no words would be split into parts), but then the lexicon would necessarily be very large, since it would have to contain all distinct words that occur in the corpus. Thus, the optimal solution needs to be a compromise between these two extremes.

Among others, de Marcken [1996], Brent [1999], Goldsmith [2001], Creutz and Lagus [2002; 2004; 2005a; 2007], and Creutz [2003; 2006] have shown that models based on the above approach produce segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle [Rissanen 1989]. Similarly, Goldwater et al. [2006] use a hierarchical Dirichlet model in combination with morph bigram probabilities.

An alternative popular approach to the segmentation of words and phrases is based on the works by Zellig S. Harris [1955; 1967]. For instance, Schone and Jurafsky [2000; 2001] make use of a Harrisian approach to suggest word stems and suffixes. In this approach, word or morpheme boundaries are proposed at locations where the predictability of the next letter in a letter sequence is low. Such a model does not use compactness of representation as an explicit optimization criterion.

Next, the so-called *Morfessor Baseline* model will be outlined; this is the principal model used in the speech recognition experiments of the current paper. Note that merely the underlying ideas and characteristics of the Morfessor model are presented; in order to find the exact mathematical and algorithmic formulation it is necessary to read previous work, e.g., Chapter 3 of Creutz [2006].

2.1 Morfessor Baseline

The Morfessor Baseline algorithm was originally introduced in Creutz and Lagus [2002], where it was called the "Recursive MDL" method. Additionally, the Baseline algorithm is described in Creutz and Lagus [2005b] and Hirsimäki et al. [2006]. The implementing computer program is publicly available for download at `http://www.cis.hut.fi/projects/morpho/`.

In slightly simplified form, the optimization criterion utilized in Morfessor Baseline corresponds to the maximization of the following posterior probability:

$$P(\text{lexicon} \mid \text{corpus}) \;\propto\; P(\text{lexicon})P(\text{corpus} \mid \text{lexicon}) = \prod_{\text{letters } \alpha} P(\alpha) \cdot \prod_{\text{morphs } \mu} P(\mu). \tag{1}$$

The lexicon consists of all distinct morphs spelled out; this forms a long string of letters α, in which each morph is separated from the next morph using a morph boundary character, which is also part of the alphabet. The probability of the lexicon is the product of the probability of each letter in this string. Analogously, the corpus is represented as a sequence of morphs, which corresponds to a particular segmentation of the words in the corpus. The probability of this segmentation equals the product of the probability of each morph token μ. Letter and morph probabilities are maximum likelihood estimates based on the corpus being modeled (empirical Bayes).

2.2 Some Remarks

Different morph segmentations are obtained if the algorithm is trained on a collection of *word tokens* vs. *word types*. The former corresponds to a *corpus*, a piece of text, where words can occur many times. The latter corresponds to a *corpus vocabulary*, where only one occurrence of every distinct word form in the corpus has been listed. As previously shown in experiments on Finnish and English text data [Creutz and Lagus 2005b], the use of a corpus vocabulary instead of the corpus itself produces segmentations that are closer

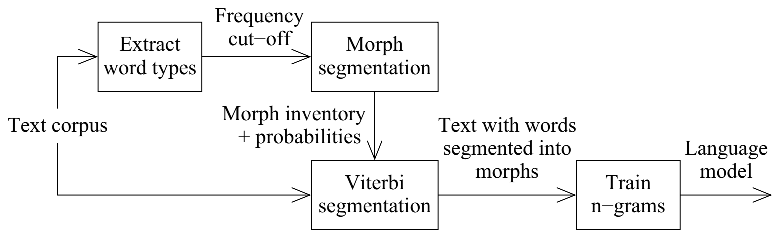


Fig. 1. (adapted from Hirsimäki et al. [2006]). How to train a segmentation model using the Morfessor Baseline algorithm, and how to further train an n-gram model based on Morfessor morphs.

to linguistic morpheme segmentations. Therefore, corpus vocabularies have been used as input to the Morfessor algorithm in the current paper.

As a result of the probabilistic (or equivalently: MDL) approach, the morph inventory discovered by the Morfessor Baseline algorithm is larger the more training data there is. In some speech recognition experiments, however, it has been desirable to restrict the size of the morph inventory. This has been achieved by setting a frequency threshold on the words on which Morfessor is trained, such that the rarest words will not affect the learning process. Nonetheless, also the rarest words can be split into morphs in accordance with the model learned, by using the Viterbi algorithm to select the most likely segmentation. The process is depicted in Figure 1.

The name of the algorithm, the Morfessor *Baseline*, was given retroactively, when more developed versions were introduced: Morfessor *Categories-ML* [Creutz and Lagus 2004] and Morfessor *Categories-MAP* [Creutz and Lagus 2005a]. The latter versions depend on the Baseline algorithm to produce an initial, or baseline, segmentation, which is then further processed. Some findings on the use of the Morfessor Categories models in speech recognition are discussed in Section 4.

3. EXPERIMENTS AND ANALYSIS

The goal of the conducted experiments is to compare n-gram language models based on morphs to standard word n-gram models in automatic speech recognition across languages. The morphs have been learned in an unsupervised manner from the LM training data using the Morfessor Baseline algorithm, which was outlined above.

3.1 Data Sets and Recognition Systems

The results from eight different tests have been analyzed. Some central properties of the test configurations are shown in Table I. The Finnish, Estonian, and Turkish test configurations correspond exactly to, or are slight variations of, experiments reported earlier in Hirsimäki et al. [2006] (Fin1: 'News task', Fin2: 'Book task'), Kurimo et al. [2006a], Kurimo et al. [2006b] (Fin3, Tur1), and Kurimo et al. [2006] (Fin4, Est, Tur2).

Three different recognition platforms have been used, all of which are state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems. The Finnish and Estonian experiments have been run on the HUT speech recognition system (Fin1 & Fin2 on a previous version described in Hirsimäki et al. [2006], and Fin3, Fin4, and Est on a newer version described in Pylkkönen [2005]). The HUT recognizer has been developed at Helsinki University of Technology. The Turkish tests were performed using the AT&T decoder [Mohri and Riley 2002]; the acoustic features were produced using the HTK front end [Young et al. 2002]. The experiments on Egyptian Colloquial Arabic were carried out using the SRI Decipher^{TM} speech recognition system (Arabic setup described in Kirchhoff et al. [2006]).

3.1.1 *Speech Data and Acoustic Models.* The type and amount of speech data vary across languages. The Finnish data consists of news broadcasts read by one single female speaker (Fin1), as well as an audio book read by another female speaker (Fin2, Fin3, Fin4). The news data and book data have been partitioned into training and test sets. The Finnish acoustic models are speaker dependent (SD). Monophones (mon) were used in the earlier experiments (Fin1, Fin2), but these were later replaced by cross-context triphones (tri).

The amount of Estonian speech data is an order of magnitude larger than the amount of Finnish data. The data consists of sentences from newspapers as well as names and digits read aloud. The training set contains utterances collected from more than 1000 speakers, and the test set has 50 speakers. Speaker-independent triphones (SI tri) have been used as acoustic models, and unsupervised speaker adaptation takes place using Cepstral Mean Subtraction (CMS) and Constrained Maximum Likelihood Linear Regression (cMLLR).

Also the Turkish data used for acoustic training contains speech from hundreds of speakers. The utterances consist of phonetically balanced sentences selected from a text corpus. The test data set is composed of newspaper sentences read by one female speaker; the same test set is used in both Turkish experiments. The acoustic models are speaker-independent triphones.

The Finnish, Estonian and Turkish data sets contain planned speech, i.e., written text read aloud. By contrast, the Arabic data consists of recorded spontaneous telephone conversations (which were only transcribed into written form afterwards). In contrast to planned speech, spontaneous speech is characterized by a large number of disfluencies: false starts and corrections, repeated words, word fragments, and so on. The speech also contains plenty of "non-speech", such as laugh and cough sounds. The Arabic data consists of the LDC CallHome corpus of Egyptian Colloquial Arabic.[1] The training data comprises the so-called 'train', 'eval96' and 'h5_new' data sets. For testing, a subset of the so-called 'dev' set was used (the rest of the 'dev' set was used as a development test set). Both the training and test sets have multiple speakers, and online speaker adaptation has been performed (as described in Kirchhoff et al. [2006]).

3.1.2 *Text Data and Language Models.* The n-gram models are trained using the SRILM toolkit [Stolcke 2002] (Fin1, Fin2, Tur1, Tur2, ECA) or corresponding software developed at HUT [Siivola and Pellom 2005] (Fin3, Fin4, Est). All models utilize the Modified Interpolated Kneser-Ney smoothing technique [Kneser and Ney 1995; Chen and Goodman 1999]. The language models are trained on text corpora much larger than the spoken texts used for acoustic training. The Arabic material is an exception, since it consists of the same corpus that was used for acoustic training. The Arabic LM training set is therefore regrettably small (160 000 words), but it does match the test set well in style, as it consists of transcribed spontaneous speech. The LM training corpora used for the other languages range in size from 17 million (Tur1) to 150 million words (Fin4). The corpora contain mainly news and book texts, which should conceivably match the style of the spoken (read)

[1] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S45`

Table I. Test configurations

	Fin1	Fin2	Fin3	Fin4	Est	Tur1	Tur2	ECA
Recognizer	HUT	HUT	HUT	HUT	HUT	AT&T	AT&T	SRI
Speech data								
Type of speech	read	read	read	read	read	read	read	spont.
Training set [kwords]	20	49	49	49	790	230	110	160
Speakers in training set	1	1	1	1	1300	550	250	310
Test set [kwords]	4.3	1.9	1.9	1.9	3.7	7.0	7.0	16
Speakers in test set	1	1	1	1	50	1	1	57
Text data								
LM training set [Mwords]	36	36	32	150	53	17	27	0.16
Models								
Acoustic models	SD mon	SD mon	SD tri	SD tri	SI tri	SI tri	SI tri	SI tri
Morph lexicon [kmorphs]	66	66	120	25	37	52	34	6.1
Word lexicon [kwords]	410	410	410	–	60	120	50	18
Morph LM size [MB]	200	200	62	820	330	160	250	0.76
Word LM size [MB]	290	290	52	–	510	150	160	1.1
Out-of-vocabulary words								
OOV LM training set [%]	5.0	5.0	5.9	–	14	5.3	9.6	0.61
OOV test set [%]	5.0	7.2	7.3	–	19	5.5	12	9.9
New words in test set [%]	2.7	3.0	3.1	1.5	3.4	1.6	1.5	9.8

test texts well.

In the morph-based models, words are split into morphs using the Morfessor Baseline algorithm, and the corpora are rewritten as sequences of morphs. Statistics are then collected for morph n-grams. As the desired output of the speech recognizer is a sequence of words rather than morphs, the recognition output is postprocessed, such that the morphs are concatenated back to words. However, in order not to lose the locations of the word boundaries, the LM explicitly needs to model which of the morph boundaries are word boundaries and which are merely word-internal boundaries. As a solution, the locations of the word boundaries are marked using a special word boundary morph, which occurs in the n-grams just like any other morph. Acoustically, the word boundary is realized as silence or not realized at all. The Arabic setup constitutes an exception also in this respect: as the SRI recognizer does not accept LM units that have no acoustic realization, word boundaries are marked on the word segments themselves; that is, each Arabic morph is context sensitive such that it is either word-final or non-word-final.

The number of distinct morphs (morph types) that constitute the lexicon varies across experiments: from 6100 (ECA) to 120 000 (Fin3); see Table I. Our previous investigations [Hirsimäki et al. 2006; Kurimo et al. 2006b] have shown that the size of the morph lexicon (within reasonable limits) does not have a large effect on the results, as long as the order of the n-grams is sufficient. For instance, there are two very similar Finnish test configurations with very different-sized morph lexicons: Fin3 (120 000) vs. Fin4 (25 000). The crucial aspect about all morph models is that a very large word vocabulary can be modeled by the concatenation of a limited set of morphs.

For comparison, standard word n-gram models have been tested. The vocabulary cannot typically include every word form occurring in the training set (because of the large number of different words), so the most frequent words are given priority; the actual lexicon sizes used in each experiment are shown in Table I. Any word not contained in the lexicon is replaced by a special out-of-vocabulary symbol. The Finnish word lexicons are large: 410 000 words. (There is no comparable word setup for Fin4.) The Tur1 lexicon contains

120 000 words, which seems sufficient. In retrospect, the word lexicons used in the Estonian and second Turkish experiment (Tur2) appear rather small (60 000 and 50 000 words, respectively). Also the Arabic word vocabulary is small (18 000 words), but this includes virtually the entire vocabulary of the LM training data.

As words and morphs are units of different length, their optimal performance may occur at different orders of the n-gram. In the experiments, the best order of the n-gram has been optimized on development test sets in the following cases: Fin1, Fin2, ECA (4-grams for both morphs and words), Tur1 (4-grams for morphs, 3-grams for words), and Tur2 (5-grams for morphs, 3-grams for words). The models have additionally been pruned using entropy-based pruning (Tur1, Tur2, ECA) [Stolcke 1998]. A different approach was used in the other experiments (Fin3, Fin4, Est) in that no fixed maximum value of n was selected. Instead, n-gram growing was performed [Siivola and Pellom 2005], such that one starts from unigrams and gradually adds those n-grams that maximize the training set likelihood. However, the size of the model is taken into account using an MDL-type complexity term, such that an n-gram is not included in the model if the resulting increase of the likelihood of the data is outweighed by the increased model size. The resulting highest order of n-grams ending up in the models was 7 for Finnish, and 8 for Estonian.

Note that the optimization procedure is neutral with respect to morphs vs. words. Table I lists the total sizes of the the language models obtained (“Morph LM size” vs. “Word LM size”). The figures are the sizes of the LM files on disk, which corresponds closely to the memory requirements for these models.[2] The LMs have been optimized with respect to performance in recognition tests (within some memory limitations); therefore, the sizes of a morph model and the corresponding word model are not necessarily the same. The Tur1 LMs are the closest in size: 160 million bytes for morphs and 150 million bytes for words. In Fin1, Fin2, Est, and ECA, the morph models are smaller than the corresponding word models. In Fin3 and Tur2, the word LMs are smaller than the morph LMs.

In terms of time requirements, it is difficult to perform a meaningful comparison of real-time factors (RTFs) across the different recognition systems (HUT, AT&T, and SRI). The Turkish experiments run in 1–2 times real time. The Finnish and Estonian tests need 8–20 times real time to complete, and the ECA experiments require a real-time factor of 5–10 (with eight processors running in parallel). When comparing the running times of systems based on morph models to the running times of word models, the morph models are typically somewhat slower. On average, the morph models take 1.5 times longer to run, most likely because of the higher branching factor in the morph LM.

3.1.3 *Out-of-Vocabulary Words.* Some statistics of the vocabulary of the training and test data have further been collected into Table I. These statistics are relevant for the assessment of the word models, as the OOV statistics define the limits of these models.

The out-of-vocabulary (OOV) rate for the LM training set corresponds to the proportion of words that were replaced by the OOV symbol in the LM training data, i.e., words that were not included in the recognition vocabulary. The high OOV rates for Estonian (14 %) and Tur2 (9.6 %) suggest that the word lexicons have poor coverage of these sets. The OOV rate for the other data sets are around 5 %, except for ECA, where practically the entire training set vocabulary has been covered.

[2] The Turkish values denote the sizes of finite state automata, and the other models have been converted to binaries from the widely used ARPA format.

Correspondingly, the test set OOV rate is the proportion of words that occur in the data sets used for running the speech recognition tests, but that are missing from the recognition lexicons. This value is thus the *minimum error* that can be obtained using the word models, or put differently, the recognizer is guaranteed to get at least this proportion of words wrong. Again, the values are high for Estonian (19 %) and Tur2 (12 %), but also for Arabic (9.9 %) because of the insufficient amount of training data. Also the other test sets suffer from fairly high OOV rates (from 5.0 to 7.3 %).

Finally, the figures labeled "new words in the test set" correspond to the proportion of words in the test set that do not occur in the LM training set. Thus, these values indicate the minimum error achievable by *any* word model trained on the training sets available (when assumed that no further data sources are available, such as dictionaries). Even if the whole training set vocabulary were included in the model, these words would still be missing from the vocabulary. A very high value is observed for Arabic (9.8 %), but the values for Finnish (Fin1, Fin2, Fin3) and Estonian are also fairly high (around 3 %).

Note that, in contrast to word models, there are no OOVs in morph models.

3.2 Results and Discussion

The overall results of the conducted speech recognition experiments are shown in Figure 2. For each experimental setup, both the result of the morph and word model are presented (with the exception of Fin4, where no comparable word experiment has been carried out).

The evaluation measure used in the upper figure (Fig. 2a) is *word accuracy* (WAC), that is, the number of correctly recognized words minus the number of incorrectly inserted words divided by the number of words in the correct reference transcription. (Another frequently used measure is the *word error rate*, WER, which relates to word accuracy as WER = 100 % – WAC.) The lower figure (Fig. 2b) displays the *letter accuracy* (LAC) for the same experiments. Letter accuracy is computed analogously to word accuracy, but with each letter in the output as its own unit. Also word breaks are considered letters in the calculation of LAC. The use of letter accuracy alongside word accuracy is motivated by a number of reasons: First, words may have different lengths in different languages and thus a comparison of word accuracies across languages is misleading. Second, very small recognition mistakes may lead to high word errors. For instance, by erroneously spelling the Finnish compound word "kaksitoistavuotias" (twelve-year-old) with a word break ("kaksitoista vuotias") nothing but errors are obtained at the word level (one substituted and one inserted word). On the letter level, by contrast, there are 18 correctly recognized and one inserted unit (the superfluous word break). Third, even though the word models are "guaranteed" to get at least the OOV words wrong, some phonetically very similar words may be suggested instead, and it may be interesting to know how large the letter errors caused by the OOV words are.

Figure 2 shows that the morph models perform better than the word models, with the exception of the Arabic experiment (ECA), where the word model outperforms the morph model. The statistical significance of these differences is confirmed by one-tailed paired Wilcoxon signed-rank tests at the significance level of 0.05.

Overall, the best performance is observed for the Finnish data sets, which is explained by the speaker-dependent acoustic models and clean noise conditions. Results are not as good for the speaker-independent models (Est, Tur1, Tur2, ECA). The Arabic setup additionally suffers from the insufficient amount of LM training data.

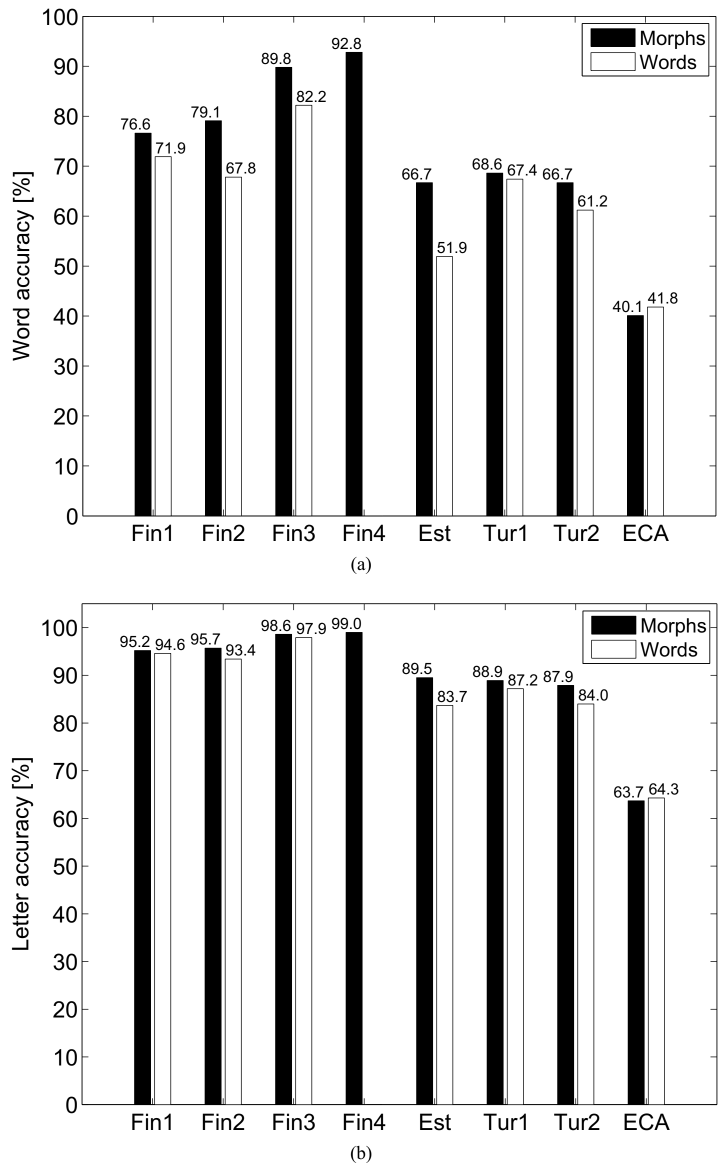


Fig. 2. Word and letter accuracies for the different speech recognition test configurations.

3.2.1 *In-Vocabulary Words.* For a further investigation of the outcome of the experiments, the test sets are partitioned into regions based on the types of words they contain. That is, the recognition output is aligned with the reference transcript and the regions aligned with *in-vocabulary* reference words (i.e., words contained in the vocabulary of the word model) are put in one partition and the remaining words (out-of-vocabulary words) are put in another partition. Word and letter accuracies are then computed separately for the two partitions. Inserted words, i.e., words that are not aligned with any word in the reference are put in the in-vocabulary partition, unless they are adjacent to an OOV region, in which case they are put in the OOV partition. This choice is motivated by the fact that these insertions are likely to be caused by the presence of the OOVs. Transitively, insertions adjacent to insertions adjacent to OOV words are also put in the OOV partition.

Figure 3 shows word and letter accuracies for the in-vocabulary words. For comparison, the overall results (from Fig. 2) are also shown as gray-shaded bars. Without exception, the accuracy for the in-vocabulary words is higher than that of the entire test set vocabulary. This is natural for the word models, since their lexicons contain these in-vocabulary words and no other words. However, also the morph models perform better on the "in-vocabulary" words, although there is no division into in-vocabulary and out-of-vocabulary words in the morph models. This occurs because the vocabularies used for the word models (the in-vocabulary words) are the most frequent words in the training data, and consequently the n-gram co-occurrence statistics for these words are the most reliable.

One could imagine that the word models would do better than the morph models on the in-vocabulary words, since the word models are focused on this subset of words, whereas the morph models reserve modeling capacity for a much larger set of words. The word accuracies in Fig. 3a also partly seem to support this view. However, Wilcoxon signed-rank tests (level 0.05) show that the difference is not statistically significant, except for Arabic and for Fin3, where the word model outperforms the morph model. Moreover, for Fin3 the superiority of the word model is statistically significant only when performance is measured in letter accuracy, not word accuracy.

With almost no exceptions, it is thus possible to draw the conclusion that morph models are capable of modeling a much larger set of words than word models *without, however, compromising the performance on the limited vocabulary covered by the word models in a statistically significant way*.

3.2.2 *Out-of-Vocabulary Words.* Following the argumentation of the previous section, suppose that a word model and morph model perform equally well on the subset of words that are included in the lexicon of the word model. However, since the morph model has better overall performance, the superiority of the morph model needs to come from its successful coping with out-of-vocabulary words. Furthermore, the larger the proportion of OOVs in the word model, the larger advantage over the word model could be expected from the morph model. This hypothesis has been investigated in Figure 4.

In Figure 4a, the OOV rate of the word model runs on the x-axis, and the y-axis shows the relative word error rate reduction achievable using a morph model instead of the word model. Each experiment (except Fin4, which lacks a word model) is situated in the plot. The relative improvement y of the morph model vs. the word model is calculated from the word error rates (WER):

$$y = 100\,\% - \frac{\text{WER(morph model)}}{\text{WER(corresponding word model)}}. \tag{2}$$

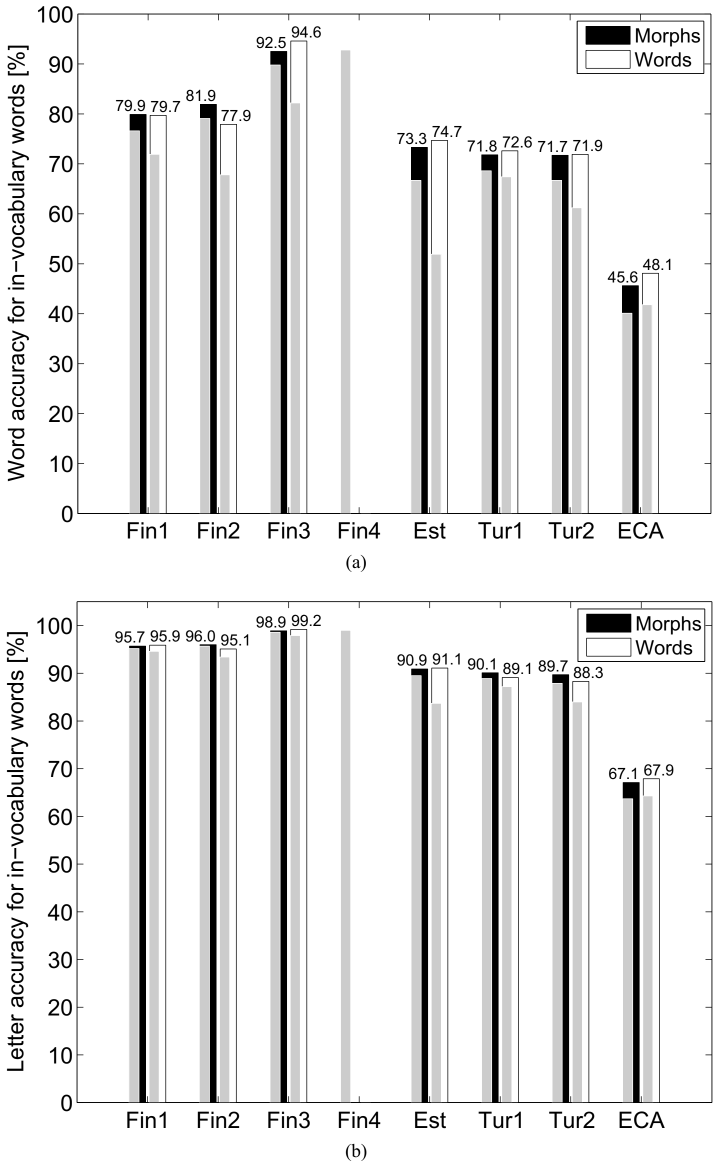
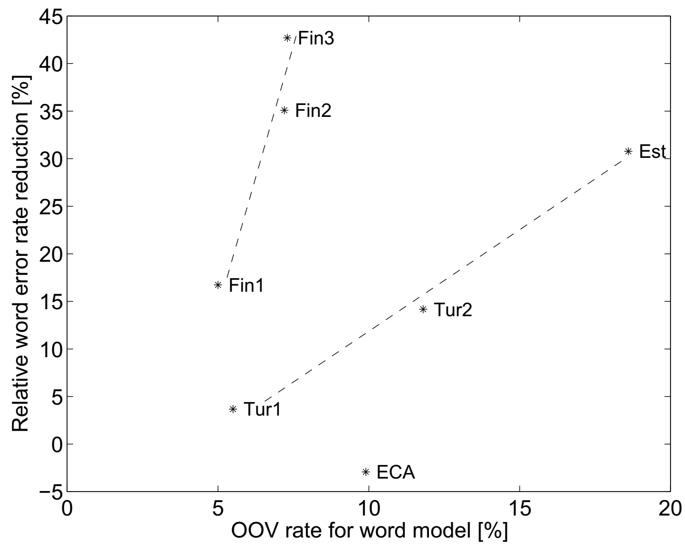
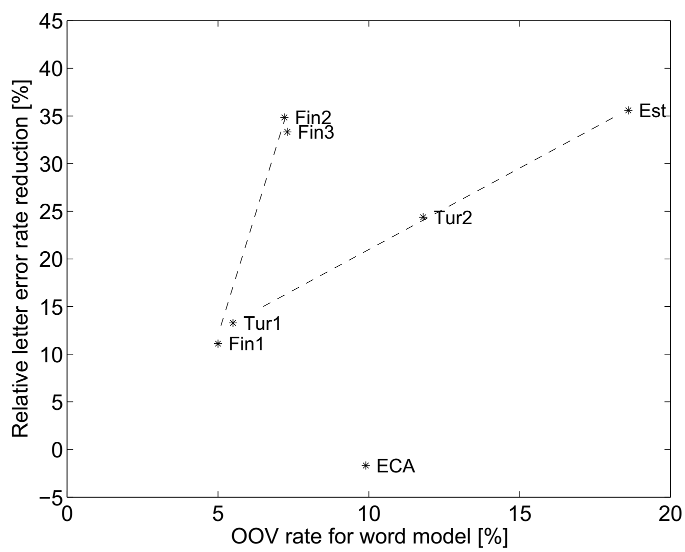


Fig. 3. Word and letter accuracies computed for only those words in the test sets that are included in the vocabularies of the word models. This corresponds to the major part of the test sets (100 % minus the figures listed on the row “OOV test set” in Table I). For comparison, the gray-shaded bars show the corresponding results for the entire test sets, i.e., all words (these results are also displayed in Figure 2).



(a)



(b)

Fig. 4. Relative word, as well as letter, error rate reduction obtained using a morph model instead of a word model, plotted against the OOV rate of the word model (i.e., the figures listed on the row "OOV test set" in Table I).

Figure 4b shows the corresponding results based on letter error rates (LER).

A look at Figures 4a and 4b does not reveal any clear correlation, across all systems, between the OOV rate of the word model and the relative advantage of the corresponding morph model. Nonetheless, if one is allowed to speculate based on such a limited number of observations, one can suggest the existence of three groups (in both the upper and lower figures): (1) in the Arabic experiment there is no correlation between the OOV rate and the advantage of the morph model, since the morph model performs slightly worse than the word model; (2) one correlation line can be drawn through the Turkish and Estonian points (3 observations); and (3) another steeper line can be drawn through the Finnish points (3 observations). The two different lines could be explained by the fact that the Turkish and Estonian systems utilize speaker-independent acoustic models, and thus the acoustic confusability is higher than with the Finnish clean speaker-dependent models. The concatenation of morphs into correct words works better under clean conditions, and the Finnish line has a steeper slope. The Estonian and Turkish systems have a lower sensitivity to acoustic subtlety, and consequently the advantage of the flexible modeling of the vocabulary is somewhat counterbalanced by mistakes that occur when putting morphs together to form words.

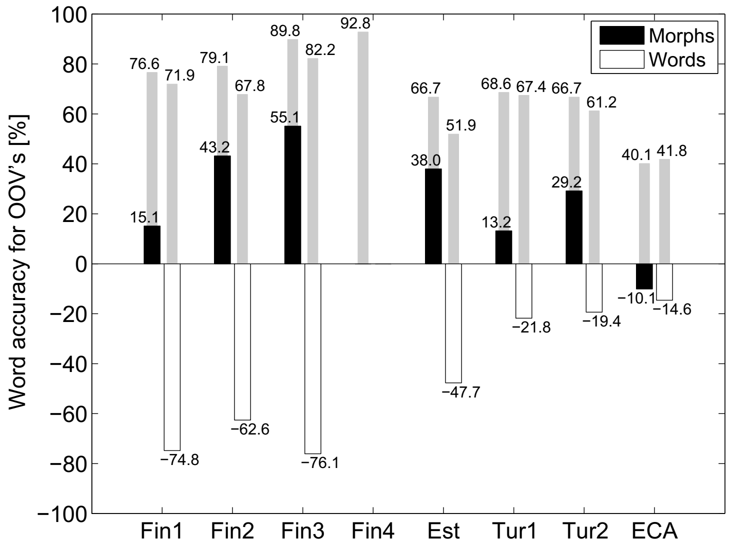
Figure 5 provides some additional information. Word and letter accuracies have been plotted for the out-of-vocabulary words contained in the test set. Thus, Figures 3 and 5 show the results separately for two partitions of the test set that together make up the entire set. It is clear from Figure 5 that the recognition accuracy for the OOVs is much lower than the overall accuracy. Also negative accuracy values are observed (Fig. 5a). This happens in situations where the number of insertions exceeds the number of correctly recognized units, which is always the case for the word models, since they are unable to recognize any OOVs correctly.

Interestingly, a correlation can again be observed within the two groups: Fin1 + Fin2 + Fin3, and Est + Tur1 + Tur2. The morph models recognize the OOVs more accurately, the higher the OOV rate is. Logically, if the OOV rate is higher, then there is more data in the OOV partition, which thus contains more frequent word forms. The presence of more data, which is less sparse, leads to better statistical estimates and thus better models.

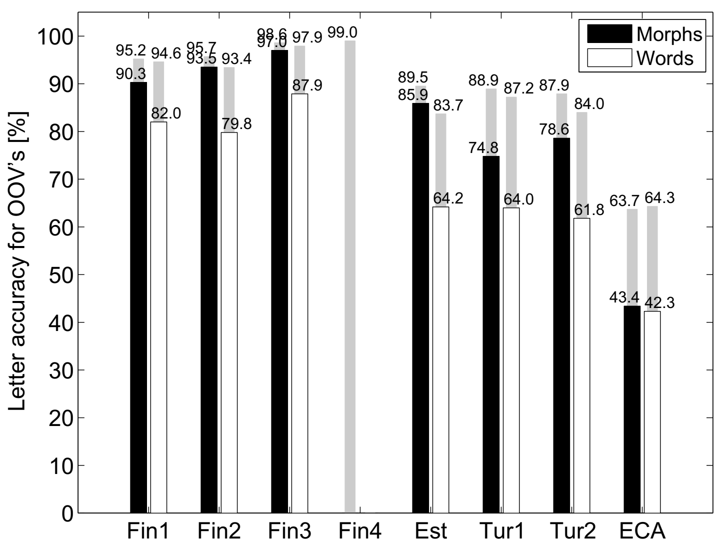
One could say that a morph model has a double advantage over a corresponding word model: the larger the proportion of OOVs in the word model is, the larger the proportion of words that the morph model can recognize but the word model cannot, a priori. In addition, the larger the proportion of OOVs, the more frequent and more "easily modelable" words are left out of the word model, and the more successfully these words are indeed learned by the morph model.

3.2.3 *New Words in the Test Set.* Since it is possible, in principle, to model any out-of-vocabulary words using morph models, it is interesting to know how good the morph models really are in a hard OOV task. The OOVs discussed above are words that were excluded from the word models, although many of them actually occurred in the training data. All words present in the training data "leave some trace" in the morph models, in the n-gram statistics that are collected for morph sequences. How, then, about new words that occur only in the test set, but not in the training set? To recognize such words correctly, the model must "invent" new combinations of morphs that it has never observed before.

Figure 6 demonstrates that the new unseen words are very challenging. Now, also the morph models mostly obtain negative word accuracies, which means that the number of in-

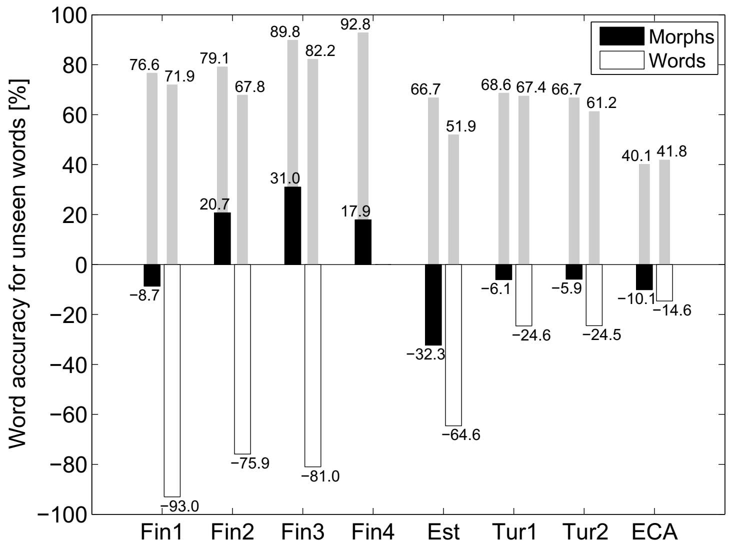


(a)

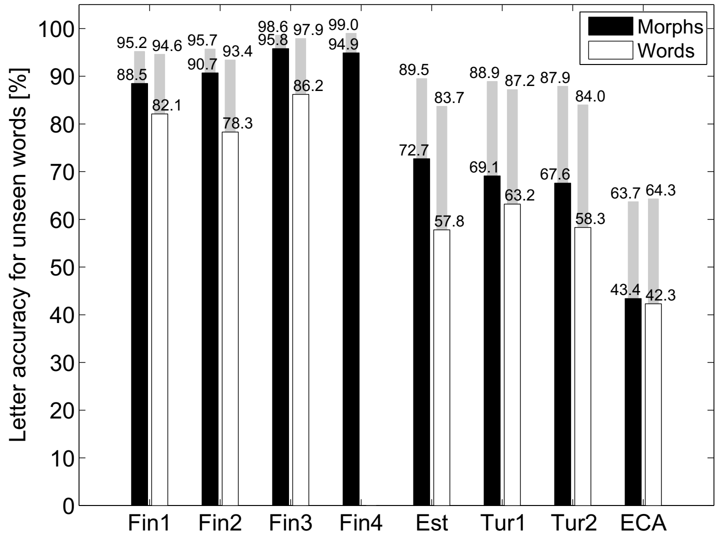


(b)

Fig. 5. Word and letter accuracies computed for those words in the test sets that are excluded from the vocabularies of the word models. The proportions of these out-of-vocabulary words in the test sets are listed on the "OOV test set" in Table I. For comparison, the gray-shaded bars show the corresponding results for the entire test sets, i.e., all words (these results are also displayed in Figure 2).



(a)



(b)

Fig. 6. Word and letter accuracies computed for the words in the test sets that do not occur at all in the training sets. This corresponds to rather small parts of the test sets (the figures listed on the row "new words in test set" in Table I). Thus, neither the word nor the morph models can learn these words directly from the training data. For comparison, the gray-shaded bars show the corresponding results for the entire test sets, i.e., all words (these results are also displayed in Figure 2).

sertions adjacent to new words exceeds the number of correctly recognized new words. For instance, the negative accuracy for Fin1 (news broadcast data) seems to be due to a number of foreign names that are difficult to get right using typical Finnish phoneme-to-grapheme mappings. However, fairly good results are obtained in the other Finnish experiments, especially Fin3. Fin3 has the largest morph lexicon (120 000 morphs) compared to the other Finnish setups (with lexicons of 25 000 and 66 000 morphs), and the new words seem to be recognized as sequences of fewer and longer morphs in Fin3 than in Fin1, Fin2, and Fin4. Apparently, this produces fewer errors.

3.3 Word Models, Vocabulary Growth, and Spontaneous Speech

To improve the word models, one could attempt to increase the vocabulary (recognition lexicon) of these models. A high coverage of the vocabulary of the training set might also reduce the OOV rate of the recognition data (test set). However, this may be difficult to obtain.

Figure 7 shows the development of the size of the training set vocabulary for growing amounts of training data. The corpora used for Finnish, Estonian, and Turkish are the data sets used for training language models (mentioned in Section 3.1.2). For comparison, a curve for English is also shown; the English corpus consists of text from the New York Times magazine. Whereas there are fewer than 200 000 different word forms in the 40 million word English corpus, the corresponding values for Finnish and Estonian corpora of the same size exceed 1.8 million and 1.5 million words, respectively. The rate of growth remains high, as the entire Finnish LM training data of 150 million words (used in Fin4) contains more than 4 million unique word forms. This value is thus ten times the size of the (rather large) word lexicon currently used in the Finnish experiments.

Figure 8 illustrates the development of the OOV rate in the test sets for growing amounts of training data. That is, assuming that the entire vocabulary of the training set is used as the recognition lexicon, the words in the test set that do not occur in the training set are OOVs. The test sets are the same as used in the speech recognition experiments, and for English, a held-out subset of the New York Times corpus was used. Again, the proportions of OOVs are fairly high for Finnish and Estonian; at 25 million words the OOV rates are 3.6 % and 4.4 %, respectively (to be compared with 1.7 % for Turkish and only 0.74 % for English). If the entire 150 million word Finnish corpus were to be used (i.e., a lexicon containing more than 4 million words), the OOV rate for the test set would still be 1.5 %.

Not surprisingly, the feasibility of the use of high-coverage standard word lexicons for Finnish and Estonian is low. In light of the plots in Figures 7 and 8, word lexicons might, however, be an option for Turkish. The slower vocabulary growth for Turkish is likely due to the much lower number of compound words in Turkish in comparison to Finnish and Estonian. Word lexicons are the state-of-the-art solution for English.

3.3.1 *Egyptian Arabic.* The vocabulary growth and OOV curves for Arabic are not visible in Figures 7 and 8, because of the small amount of Arabic data available (164 000 words). However, Figures 9 and 10 provide a "close-up" of the first 164 000 words, including Arabic. The data sets shown in Figures 7 and 8 all consist of planned, written text, whereas the ECA corpus contains unplanned, transcribed spontaneous speech. Because of these differences, the type of text (planned or spontaneous) has been indicated explicitly in the new figures.

Additional sources have been provided for Arabic and English: planned Arabic text

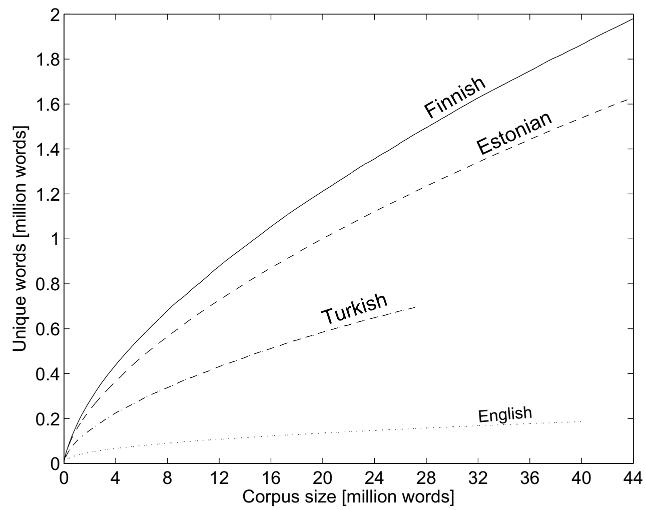


Fig. 7. Vocabulary growth curves for different languages: For growing amounts of text (word tokens) the numbers of unique different word forms (word types) occurring in the text are plotted.

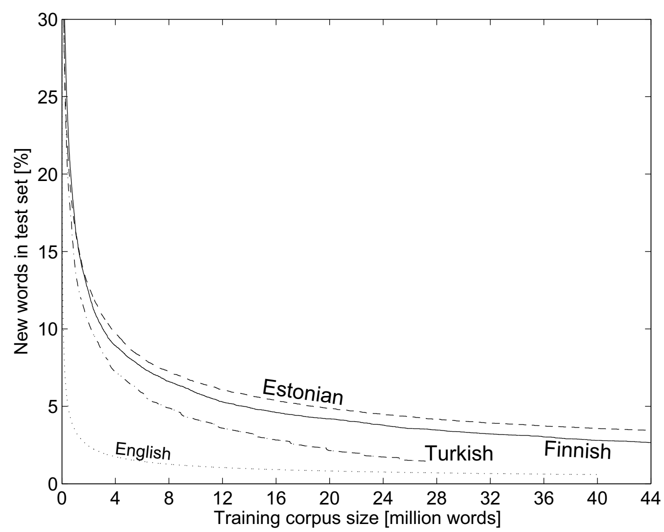


Fig. 8. For growing amounts of training data: development of the proportions of words in the test set that are not covered by the training set.

from the FBIS corpus of Modern Standard Arabic (a collection of transcribed radio newscasts from various radio stations in the Arabic-speaking world), as well as spontaneous transcribed English telephone conversations from the Fisher corpus.[3] The point here is to illustrate that a smaller, more slowly growing vocabulary is used in spontaneous speech than in planned speech.

The smaller vocabulary also appears to result in lower OOV rates for spontaneous speech compared to planned speech; this can be verified by comparing the two English curves in Figure 10. It is true that the Arabic 'planned' curve descends almost to the same level as the 'spontaneous' curve in Figure 10, but this probably occurs because short vowels are not marked in the FBIS corpus (planned Arabic). The lack of short vowels in written Arabic increases ambiguity, since several word forms may be mapped onto the same written form. Therefore, it is without doubt that the vocabulary growth curve for planned Arabic in Figure 9 would rise more steeply if vowels were marked in the script, and the gap between planned and spontaneous Arabic would increase. (All short vowels *are* marked in the ECA corpus.) It is also likely that the OOV curve for planned Arabic would run at a higher level in Figure 10, if the vowels were present.

What the Arabic 'spontaneous' curves reveal is that they are located fairly close to the English 'planned' curves and much below the Finnish, Estonian, and Turkish curves. Thus, even though Arabic is considered a "morphologically rich" language, this is not manifested through a considerable vocabulary growth (and high OOV rate) in the Egyptian Colloquial Arabic data used in the current speech recognition experiments. Consequently, it may not be that surprising that the morph model did not work particularly well for Arabic.

The Arabic experiment is problematic in the sense that the amount of LM training data is much too small, and the language models trained on this data, be they word or morph based, are poor. Obtaining more suitable data is hard, since Egyptian Colloquial Arabic is a spoken dialect that is very rarely written.

Arabic words consist of a stem surrounded by prefixes and suffixes, which are fairly successfully segmented out by Morfessor. However, Arabic also has *templatic* morphology, in which the word stem is formed through the insertion of a vowel pattern into a "consonantal skeleton"; e.g., the consonant sequence 'k-t-b' means 'writing-related', and the following stems can be formed, among others: 'kitaab' (book), 'kutub' (books), 'kaatib' (writer). In this sense, Arabic word structure differs from the agglutinative morphology of Finnish, Estonian, and Turkish, where morphemes are "glued" together one after another. Note, however, that word forming in Finnish, Estonian, and Turkish is not purely agglutinative. Morpho-phonological phenomena exist that alter morphemes depending on their context, which makes the task more challenging for Morfessor.

Additional experiments have been performed with the Arabic data and Factored Language Models (FLMs) [Kirchhoff et al. 2006], using exactly the same setup as in the Morfessor experiments. The FLM is a powerful model that makes use of several sources of information, in particular a morphological lexicon of ECA. The FLM also incorporates the special traits of the Arabic templatic morphology. Despite its sophistication, the FLM barely outperforms the standard word model: The word accuracy of the FLM is 42.3 % and that of the word model is 41.8 %; the letter accuracies are 64.6 % and 64.3 %, respectively. These differences are statistically significant. (Both the FLM and word model outperform the morph model, which obtains a word accuracy of 40.1 % and a letter accuracy of 63.7 %,

[3] Available at `http://www.ldc.upenn.edu/`.

differences that are also statistically significant.) The speech recognition implementation of both the FLM and the word model is based on *whole words* (although subword units are used for assigning probabilities to word forms in the FLM). This contrasts these models with the morph model, which splits words into subword units also in the speech recognition implementation. It seems that the splitting is a source of errors in this "fragile" experimental setup with very little data available.

4. FURTHER DISCUSSION

Since the use of subword units generated by the Morfessor Baseline algorithm has been shown to improve speech recognition in many cases, one could also explore alternative subword-based methods. In the following, some additional work will be discussed briefly, and parallels to other previous work will be drawn.

4.1 Segmentations Corresponding Closer to Linguistic Morpheme Segmentations

When segmentations produced by the Morfessor Baseline method are compared to linguistic morpheme segmentations, the algorithm suffers from three types of fairly common errors: *undersegmentation* of frequent strings, *oversegmentation* of rare strings, and *morphotactic violations*. This follows from the fact that the most concise representation (according to the optimization criterion) is obtained when any frequent string is stored as a whole in the morph lexicon (e.g., English 'having, soldiers, states, seemed'), whereas infrequent strings are better coded in parts (e.g., 'or+p+han, s+ed+it+ious, vol+can+o'). Morphotactic violations are a consequence of the context-independent nature of the model: For instance, the morphs '-s' and '-ed' are frequently occurring *suffixes* in the English language, but the context-insensitive algorithm occasionally suggests them in word-initial position as *prefixes* ('s+wing, ed+ward, s+urge+on').

4.1.1 *The Context-Sensitive Morfessor Categories-ML and Morfessor Categories-MAP Model Versions Applied in Speech Recognition.* Morfessor Categories-ML [Creutz and Lagus 2004] introduces morph categories. The segmentation of the corpus is modeled using a Hidden Markov Model (HMM) with transition probabilities between categories and emission probabilities of morphs from categories. Three categories are used: *prefix, stem*, and *suffix* and an additional *non-morpheme* (or *noise*) category. Some distributional properties of the morphs in a proposed segmentation of the corpus are used for determining category-to-morph emission probabilities. A morph that is observed to precede a large number of different morphs is a likely prefix (e.g., English 're-, un-, mis-'). Correspondingly, a morph that is observed to follow a large set of morphs is likely to be a suffix (e.g., '-s, -ed, -ing'). A morph that is not very short is likely to be a stem (e.g., 'friend, hannibal, poison'). A morph that is not an obvious prefix, stem, or suffix in the position it occurs may be an indication of an erroneous segmentation. Such morphs are tagged as noise (e.g., all morphs in the segmentation 'vol+can+o').

The identification of "noise" and likely erroneous segmentations makes it possible to apply some heuristics in order to partly remedy the shortcomings of Morfessor Baseline. Undersegmentation is reduced by forcing splits of redundant morphs in the morph lexicon. These morphs consist of other morphs that are also present in the lexicon (e.g., 'seemed = seem+ed'). Some restrictions apply, such that splitting into noise morphs is prohibited. The opposite problem, oversegmentation, is alleviated by joining morphs tagged as noise

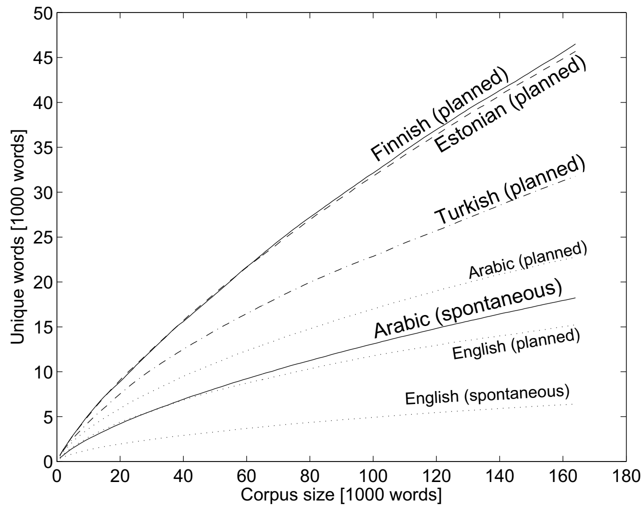


Fig. 9. Vocabulary growth curves for the different languages, including Arabic: For growing amounts of text (word tokens) the numbers of unique different word forms (word types) occurring in the text are plotted. Each data set is marked as either *planned* written text or unplanned transcribed *spontaneous* speech.

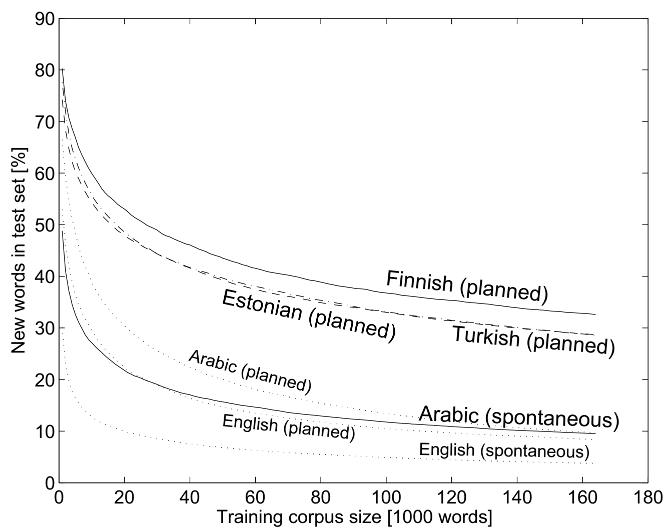


Fig. 10. For growing amounts of training data: development of the proportions of words in the test set that are not covered by the training set.

with their neighbors (e.g, ‘vol+can+o’ becomes ‘volcano’). Morphotactic violations are less likely to occur because of the context sensitivity of the HMM model.

As the next step in the development, the Morfessor Categories-MAP model version [Creutz and Lagus 2005a] emerged in an attempt to reformulate Categories-ML in a more elegant fashion. In Categories-ML, the optimal segmentation of the corpus is sought through Maximum Likelihood (ML) reestimation, whereas the complexity of the morph lexicon is controlled heuristically. In a Maximum a Posteriori (MAP) model, an explicit probability is calculated for both the lexicon and the representation of the corpus conditioned on the lexicon. Categories-MAP and the Baseline method are MAP models.

The Morfessor Categories-ML and Categories-MAP model versions clearly outperform the Morfessor Baseline algorithm when it comes to a *linguistic* morpheme segmentation task. The segmentations proposed by the Categories models match grammatical morpheme segmentations more accurately than the segmentations produced by the Baseline model; results for Finnish, English, Turkish, and Egyptian Colloquial Arabic can be found in Creutz [2006].

However, no clear advantage over the Baseline method has so far been observed in speech recognition, when n-gram language models have been estimated for words split into morphs according to the Morfessor Categories models. Speech recognition results are available for the experimental setups Fin3, Tur1, and ECA. In Fin3 no statistically significant differences could be found between the Morfessor models. In Tur1, the difference between the Baseline and Categories-MAP is not statistically significant, whereas the Categories-ML model is slightly worse (statistically significantly only when measured in word accuracy, not letter accuracy). In ECA, Categories-ML and Categories-MAP are slightly better than the Baseline, but the difference is statistically significant only when measured in letter accuracy, not word accuracy.

4.1.2 *Linguistic Gold Standard Segmentations.* When linguistically correct segmentations of words into morphemes exist, one can use this “gold standard” as a basis for the n-gram language models. That is, the morphs are based on linguistic knowledge rather than unsupervised methods such as Morfessor. Drawbacks of the linguistic approach are their labor-intensive design and the fact that all manually made systems suffer from out-of-vocabulary words to some degree, i.e., words that have not been coded into the system. A potentially important strength, however, is the high accuracy of such systems based on expert knowledge. Nonetheless, no statistically significant difference between gold-standard morphs and Morfessor Baseline morphs have been observed in the tested speech recognition setups (Fin3, Tur1, ECA), with the exception of the slightly better performance of the ECA gold standard (measured in letter accuracy only, not word accuracy).

4.2 Better Modeling of OOVs in Word Models

Instead of using subword units throughout the entire recognition system (as has been advocated above), one can augment a standard word model with specialized models for coping with OOVs. The goal of OOV modeling may be to actually recognize and spell new words as accurately as possible, or one may content oneself with the *detection* of OOVs and mark them using a special OOV symbol. According to a rule of thumb, each OOV results in 1.5 to 2 errors on average [Klakow et al. 1999; Bisani and Ney 2005], since words adjacent to an OOV are typically misrecognized because of the presence of the OOV. Therefore, the successful detection of OOVs may reduce the error rate for in-vocabulary words.

For instance, Gallwitz et al. [1996] make use of special HMMs for OOVs in their experiments on German. The word error rate on the in-vocabulary words decreases by 6 % (relative), even though the OOV detection rate is fairly low (15 % correct). Klakow et al. [1999] construct multi-phoneme word fragments to be used as fillers in an English continuous speech recognition task; the fillers reduce the damage on in-vocabulary words (by 24 % relative) as they detect OOV regions (47 % correct) and provide a phonetic transcription for these regions. Bazzi and Glass [2000] augment an English word-based recognizer with a phone-based OOV model, so that any OOV word can be realized as an arbitrary sequence of phones. They detect half of the OOVs correctly, but observe a small degradation in the overall word error rate. In later work, Bazzi and Glass [2001; 2002] experiment with variable-length units as well as multi-class OOV models, and are able to improve the OOV detection rate to 70 %. The error rate for in-vocabulary words is still slightly higher (by 3 % relative) than that of the baseline system. Mou and Zue [2001] model the sub-lexical structures of OOV words and obtain improvements in the language model perplexity on a held-out test set; however, no speech recognition tests were performed. Galescu [2003] does perform speech recognition experiments and obtains small improvements (0.7 % and 1.9 % relative word error rate reduction) on English news broadcast data, when he augments a 20 000-word model with 11 000 automatically derived subword units. Bisani and Ney [2005] perform similar experiments using shorter subword units but higher-order n-grams in a Wall Street Journal dictation task. Their model also improves consistently over word baselines of different sizes (5 000, 20 000, and 64 000 words); the improvement is less prominent the larger the baseline lexicon (0.6 % relative word error rate reduction for 64 000 words).

In the experimental setups Fin1, Fin2, Fin3, Fin4, and Tur1, word models augmented with phonemes have been tested. In these models, the OOVs in the LM training data have been split into phonemes, and n-gram statistics have been computed over sequences of in-vocabulary words with interspersed sequences of phonemes. Thus, ideally, this hybrid word model can recognize OOVs in the test set by concatenating phonemes. As a result, in Fin1, Fin2, and Fin3, the word model augmented with phonemes performs better than the standard word model, but worse than the morph model. In Fin4, no standard word model is available, but a 400 000-word lexicon augmented with phonemes obtains slightly worse results than the morph model. In Tur1, words augmented with phonemes perform on a par with the standard word model. However, the differences are small: In Fin1 and Fin2 the difference between the two word models is statistically significant only when measured in word accuracy, whereas the difference between the morph model and the word models is statistically significant for both word and letter accuracy. In Fin3, interestingly, the difference between the morph model and the phoneme-augmented word model is not statistically significant, while the difference between these models and the standard word model is. Also in Fin4, the difference between the morph model and the phoneme-augmented word model is not statistically significant. In Tur1, all three models are very close; statistically significant differences exist between the morph model and the word models but not between the word models. For all experimental setups, on the partition of OOVs only, the phoneme-augmented word models perform better than the standard word models but worse than the morph models; not all differences are statistically significant.

Taken together, hybrid phoneme-augmented word models may be an alternative to morph models, at least if the acoustic data is clean (speaker-dependent triphones as in Fin3 and

Fin4) and the amount of LM data available for training OOV phoneme sequences is large (Fin4). When the acoustic models are less precise (monophones in Fin1 and Fin2, speaker-independent triphones in Tur1), morph models seem to be better.

4.3 Grapheme-to-Phoneme Mapping

The mapping between graphemes (letters) and phonemes is straightforward in the languages studied in the current paper. More or less, there is a one-to-one correspondence between letters and phonemes. That is, the spelling of a word indicates the pronunciation of the word, and when splitting the word into parts, the pronunciation of the parts in isolation does not differ much from the pronunciation of the parts in context.

In reality, however, one-to-one grapheme-to-phoneme mapping is an oversimplification, since many loan words do not comply with the phonetic spelling rules of the language. In addition, especially in spontaneous speech, contractions are frequent; e.g., the Finnish word ‘kaksikymmentäkuusi’ (twenty-six) is usually pronounced ‘kakskytkuus’. In spite of this, simple letter-to-phoneme conversions have been used in the experiments on Finnish, Estonian, and Turkish.

A more advanced technique was used for Egyptian Colloquial Arabic: a spelled word was split using Morfessor, and a segmentation of the pronunciation of the word was obtained by maximum-likelihood alignment of the characters in both strings (spelling vs. pronunciation) and inserting breaks into the pronunciation at the locations given by the alignment. In cases where one spelled morph got different pronunciations in different contexts, the different variants were made unique through numbering. The n-gram LM could then learn which variant to use in which context. For example, in English this would correspond to having two versions of the morph ‘hid’ (in ‘hid’ vs. ‘hiding’): ‘hid1’ vs. ‘hid2 + ing’. In future work, this technique could be tested also on the other languages. Note that the ECA script used in the experiments was not ordinary Arabic script; first, the Egyptian dialect is seldom written; second, when Arabic is written it is customary to use Arabic rather than Latin letters; third, short vowels are usually omitted from written Arabic.

5. CONCLUSIONS

It has been confirmed in this paper that morph-based language models systematically outperform standard word-based models in a number of speech recognition experiments on Finnish, Estonian, and Turkish data. The superiority of the morph models is due to their better handling of out-of-vocabulary words. Importantly, the better performance on OOVs does not degrade performance on in-vocabulary words in comparison to word models in a statistically significant way.

In experiments on Egyptian Colloquial Arabic, the morph-based language model was outperformed by the standard word-based model. The Arabic data was the only data set consisting of spontaneous rather than planned speech, and the morphology was not “as rich” as one could have expected; this was manifested through a slower vocabulary growth rate, almost on par with “morphologically poor” written English. In addition, the data available for training the Arabic language model was lamentably small, which led to very high error rates for all tested models.

Hybrid phoneme-augmented word models that are capable of modeling OOVs as arbitrary sequences of phonemes did not outperform the morph models on the tested languages (Finnish and Turkish). Among different morph models, the simple Morfessor Baseline model was not outperformed by more linguistically motivated segmentation methods (the Morfessor Categories models and grammatical gold-standard morpheme segmentations). However, to do justice to these more advanced models one might consider using Factored Language Models (which worked relatively well in the Arabic experiments) or some other refined techniques as a complement to the currently used standard n-gram LMs.

ACKNOWLEDGMENTS

The authors express their sincere gratitude to Katrin Kirchhoff and Dimitra Vergyri for their valuable assistance related to the Arabic experiments. We are grateful to the anonymous reviewers for insightful comments and to the SRI editor for proofreading the text. We would like to thank the EU AMI training program for funding part of this work. The work was also partly funded by DARPA under contract No. HR0011-06-C-0023 (approved for public release, distribution is unlimited). The views herein are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

ARISOY, E., DUTAĞACI, H., AND ARSLAN, L. M. 2006. A unified language model for large vocabulary continuous speech recognition of Turkish. *Signal Processing 86,* 10 (October), 2844–2862.

ARISOY, E. AND SARAÇLAR, M. 2006. Lattice extension and rescoring based approaches for LVCSR of turkish. In *Proceedings of Interspeech – ICSLP'06*. 1025–1028.

BAZZI, I. AND GLASS, J. 2001. Learning units for domain-independent out-of-vocabulary word modelling. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*. Aalborg, Denmark.

BAZZI, I. AND GLASS, J. R. 2000. Modeling out-of-vocabulary words for robust speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. Beijing, China.

BAZZI, I. AND GLASS, J. R. 2002. A multi-class approach for modelling out-of-vocabulary words. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*. Denver, CO, USA.

BERTON, A., FETTER, P., AND REGEL-BRIETZMANN, P. 1996. Compound words in large-vocabulary German speech recognition systems. In *Proceedings of ICSLP '96*. Vol. 2. Philadelphia, PA, USA, 1165–1168.

BILMES, J. A. AND KIRCHHOFF, K. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the Human Language Technology Conference (HLT/NAACL)*. Edmonton, Canada, 4–6.

BISANI, M. AND NEY, H. 2005. Open vocabulary speech recognition with flat hybrid models. In *Proceedings of Interspeech'05*. Lisbon, Portugal.

BRENT, M. R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning 34*, 71–105.

BYRNE, W., HAJIČ, J., IRCING, P., JELINEK, F., KHUDANPUR, S., KRBEC, P., AND PSUTKA, J. 2001. On large vocabulary continuous speech recognition of highly inflectional language — Czech. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*. 487–489.

CHEN, S. F. AND GOODMAN, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language 13*, 359–394.

CREUTZ, M. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. 280–287.

CREUTZ, M. 2006. Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition. Ph.D. thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology. `http://lib.tkk.fi/Diss/2006/isbn9512282119/`.

CREUTZ, M. AND LAGUS, K. 2002. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL'02*. Philadelphia, Pennsylvania, USA, 21–30.

CREUTZ, M. AND LAGUS, K. 2004. Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*. Barcelona, 43–51.

CREUTZ, M. AND LAGUS, K. 2005a. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*. Espoo, Finland, 106–113.

CREUTZ, M. AND LAGUS, K. 2005b. 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.

CREUTZ, M. AND LAGUS, K. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing 4,* 1 (January).

DE MARCKEN, C. G. 1996. Unsupervised language acquisition. Ph.D. thesis, MIT.

GALESCU, L. 2003. Recognition of out-of-vocabulary words with sub-lexical language models. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*. Geneva, Switzerland, 249–252.

GALLWITZ, F., NÖTH, E., AND NIEMANN, H. 1996. A category based approach for recognition of out-of-vocabulary words. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*. Philadelphia, PA, USA.

GEUTNER, P., FINKE, M., AND SCHEYTT, P. 1998. Adaptive vocabularies for transcribing multilingual broadcast news. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2. 925–928.

GOLDSMITH, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics 27,* 2, 153–198.

GOLDWATER, S., GRIFFITHS, T. L., AND JOHNSON, M. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL)*. Sydney, Australia, 673–680.

HACIOGLU, K., PELLOM, B., CILOGLU, T., OZTURK, O., KURIMO, M., AND CREUTZ, M. 2003. On lexicon creation for Turkish LVCSR. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*. 1165–1168.

HARRIS, Z. S. 1955. From phoneme to morpheme. *Language 31,* 2, 190–222. Reprinted 1970 in *Papers in Structural and Transformational Linguistics*, Reidel Publishing Company, Dordrecht, Holland.

HARRIS, Z. S. 1967. Morpheme boundaries within words: Report on a computer test. *Transformations and Discourse Analysis Papers 73*. Reprinted 1970 in *Papers in Structural and Transformational Linguistics*, Reidel Publishing Company, Dordrecht, Holland.

HIRSIMÄKI, T., CREUTZ, M., SIIVOLA, V., KURIMO, M., VIRPIOJA, S., AND PYLKKÖNEN, J. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language 20,* 4, 515–541.

KIRCHHOFF, K., VERGYRI, D., BILMES, J., DUH, K., AND STOLCKE, A. 2006. Morphology-based language modeling for Arabic speech recognition. *Computer Speech and Language 20,* 4, 589–608.

KLAKOW, D., ROSE, G., AND AUBERT, X. 1999. OOV-detection in large vocabulary system using automatically defined word-fragments as fillers. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech)*. Budapest, Hungary, 49–52.

KNEISSLER, J. AND KLAKOW, D. 2001. Speech recognition for huge vocabularies by using optimized subword units. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*. 69–72.

KNESER, R. AND NEY, H. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*. 181–184.

KURIMO, M., CREUTZ, M., VARJOKALLIO, M., ARISOY, E., AND SARAÇLAR, M. 2006a. Unsupervised segmentation of words into morphemes – Challenge 2005, an introduction and evaluation report. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*. Venice, Italy.

KURIMO, M., CREUTZ, M., VARJOKALLIO, M., ARISOY, E., AND SARAÇLAR, M. 2006b. Unsupervised segmentation of words into morphemes – Morpho Challenge 2005, Application to automatic speech recognition. In *Proceedings of Interspeech 2006*. Pittsburgh, PA, USA.

KURIMO, M., PUURULA, A., ARISOY, E., SIIVOLA, V., HIRSIMÄKI, T., PYLKKÖNEN, J., ALUMÄE, T., AND SARAÇLAR, M. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2006*. New York, USA.

KWON, O.-W. AND PARK, J. 2003. Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication 39,* 3–4, 287–300.

LARSON, M., WILLETT, D., KOEHLER, J., AND RIGOLL, G. 2000. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proceedings of ICSLP*.

MOHRI, M. AND RILEY, M. D. 2002. DCD library, Speech recognition decoder library. AT&T Labs Research. `http://www.research.att.com/sw/tools/dcd/`.

MOU, X. AND ZUE, V. 2001. Sub-lexical modelling using a finite-state transducer framework. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Salt Lake City, UT, USA.

ORDELMAN, R., VAN HESSEN, A., AND DE JONG, F. 2003. Compound decomposition in Dutch large vocabulary speech recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*. 225–228.

PYLKKÖNEN, J. 2005. An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition. In *Proceedings of the 2nd Baltic Conference on Human Language Technologies (HLT'2005)*. 167–172.

RISSANEN, J. 1989. *Stochastic Complexity in Statistical Inquiry*. Vol. 15. World Scientific Series in Computer Science, Singapore.

SCHONE, P. AND JURAFSKY, D. 2000. Knowledge-free induction of morphology using Latent Semantic Analysis. In *Proceedings of CoNLL-2000 & LLL-2000*. 67–72.

SCHONE, P. AND JURAFSKY, D. 2001. Knowledge-free induction of inflectional morphologies. In *Proceedings of NAACL-2001*.

SHAFRAN, I. AND HALL, K. 2006. Corrective models for speech recognition of inflected languages. In *Proceedings of EMNLP*. Sydney, Australia.

SIIVOLA, V. AND PELLOM, B. 2005. Growing an n-gram model. In *Proceedings of Interspeech'05*. Lisbon, Portugal, 1309–1312.

STOLCKE, A. 1998. Entropy-based pruning of backoff language models. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. Lansdowne, VA, USA, 270–274.

STOLCKE, A. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*. 901–904. `http://www.speech.sri.com/projects/srilm/`.

WHITTAKER, E. AND WOODLAND, P. 2000. Particle-based language modelling. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. 170–173.

YOUNG, S., OLLASON, D., VALTCHEV, V., AND WOODLAND, P. 2002. *The HTK Book (for version 3.2 of HTK)*. University of Cambridge.

Received Month Year; revised Month Year; accepted Month Year