# Unsupervised segmentation of words into morphemes – Challenge 2005 An Introduction and Evaluation Report

**Mikko Kurimo, Mathias Creutz, Matti Varjokallio**
Adaptive Informatics Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland
`Mikko.Kurimo@tkk.fi`

**Ebru Arisoy and Murat Saraclar**
Bogazici University
Electrical and Electronics Eng. Dept.
34342 Bebek, Istanbul, TURKEY

## Abstract

The objective of the challenge for the unsupervised segmentation of words into morphemes, or shorter the *Morpho Challenge*, was to design a statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. Ideally, these are basic vocabulary units suitable for different tasks, such as speech and text understanding, machine translation, information retrieval, and statistical language modeling. The segmentations were evaluated in two complementary ways: *Competition 1:* The proposed morpheme segmentation were compared to a linguistic morpheme segmentation gold standard. *Competition 2:* Speech recognition experiments were performed, where statistical n-gram language models utilized the proposed word segments instead of entire words. Data sets were provided for three languages: Finnish, English, and Turkish. Participants were encouraged to apply their algorithm to all of these test languages.

## 1 Introduction

Segmentation is a common problem in the analysis of data from many modalities such as gene sequences, image analysis, time series, and segmentation of text into words. It is conceivable that similar machine learning methods could work well in different segmentation tasks.

The task proposed here was to design a statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. The purpose is to obtain a set of basic vocabulary units for different tasks, such as speech and text understanding, machine translation (Lee, 2004), information retrieval (Zieman and Bleich, 1997), and statistical language modeling (Geutner, 1995; Hirsimäki et al., 2006).

In many European languages this task is both difficult and necessary, due to the large number of different word forms found in text. In highly-inflecting languages, such as Finnish and Hungarian, there may be thousands of different word forms of the same root, which makes the construction of a fixed lexicon for any reasonable coverage hardly feasible. Also in compounding languages, such as German, Swedish, Greek and Finnish, complex concepts can be expressed in one single word, which considerably increases the number of possible word forms and calls for the use of sub-word segments as vocabulary units.

The discovery of meaningful word segments has already shown to be relevant for language modeling for speech recognition in Finnish, Turkish and Estonian (Hirsimäki et al., 2006; Kurimo et al., 2006), where language models based on statistically discovered sub-word units have rivaled language models that utilize words. However, any of the research fields dealing with natural language of any kind, as well as multimodal integration, is expected to benefit from the discovery of general meaning-bearing units. For example, a machine translator could have a vocabulary based on minimal meaningful units and generate output words and sentences using them (e.g., translation from English to Finnish: `fact+s about our car+s` / `tieto+a auto+i+sta+mme`). In information retrieval, some of the units (the word roots or stems) might be utilized as key words whereas others might be discarded (e.g., `tietoa` `autoistamme` → `tieto auto`; Engl. `fact car`).

A good segmentation algorithm should be able to find units that are meaningful (that is, usable for representing text for many different tasks), that cover as much of the naturally occurring language as possible (including unseen words), and that can be used to generate the totality of the language. The field of linguistics has attempted to capture these properties by the concept of "morpheme", the difference being that a morpheme may not correspond directly to a particular word segment but to an abstract class. However, in this challenge the task was to uncover concrete word segments.

In obtaining such a segmentation, the use of linguistic analysis and manual coding may be an option for some languages, but not all, due to being very labor-intensive. Furthermore, statistical machine learning methods might eventually discover models that rival even the most carefully linguistically designed morphologies.

In order to be a morphology-discovery method the method should be very language-general, that is, applicable to many different languages without the manual coding of language dependent rules, etc. An example of a general morphology-discovery method is described in (Creutz and Lagus, 2005a).

The main challenge in the task is the sparsity of language data: A significant portion of the words may occur only once even in the largest corpora. Thus, the algorithm should learn meaningful word segments (i.e., inner structures of words) and be capable of generalizing to previously unseen words.

## 2 Task

The task was the unsupervised segmentation of word forms into sub-word units (segments) given a data set that consists of a long list of words and their frequencies of occurrence in a corpus. The number of unique segments was restricted to the range 1000 - 300,000 (type count). Most of the participants, however, failed to keep the number of segments below 300,000, so it was decided to disregard this limitations and accept all submissions.

Data sets were provided for three languages: Finnish, English, and Turkish. Participants were encouraged to apply their algorithm to all of these test languages. Solutions, in which a large number of parameters must be ”tweaked” separately for each test language were discouraged, since the aim of the challenge was the unsupervised (or very minimally supervised) segmentation of words into morphemes. It was required that the participants submitted clear descriptions of which steps of supervision or parameter optimization were involved in the algorithms.

The segmentations were evaluated in two complementary ways: *Competition 1:* The proposed morpheme segmentation were compared to a linguistic morpheme segmentation gold standard (Creutz and Linden, 2004). *Competition 2:* Speech recognition experiments were performed, where statistical n-gram language models utilized the proposed word segments instead of entire words. Competition 1 included all three test languages. Winners were selected separately for each language. As a performance measure, the F-measure of accuracy of discovered morpheme boundaries was utilized. Should two solutions have produced the same F-measure, the one with higher precision would win. Competition 2 included speech recognition tasks in Finnish and Turkish. The organizers trained a statistical language model based on the segmentations and performed the required speech recognition experiments. As a performance measure, the phoneme error rate in speech recognition was utilized.

## 3 Data sets

The data sets provided by the organizers consisted of word lists. Each word in the list was preceded by its frequency in the corpora used. The participants' task was to return exactly the same list(s) of words, with spaces inserted at the locations of proposed morpheme boundaries.

For instance, a subset of the supplied English word list looked like this:

```
6755 sea
1 seabed
1 seabeds
2 seabird
34 seaboard
1 seaboards
```

Submission for this particular set of words might have looked like this:

```
sea
sea bed
sea bed s
sea bird
sea board
```

sea board s

The Finnish word list was extracted from newspaper text and books stored at the Language Bank of CSC [1]. Additionally, newswires from the Finnish National News Agency were used.

The English word list was based on publications and novels from the Gutenberg project, a sample of the English Gigaword corpus, as well as the entire Brown corpus.

The Turkish word list was based on prose and publications collected from the web, newspaper text, and sports news.

The desired segmentations, according to the gold standard (Creutz and Linden, 2004), for a small sample of words (500 – 700 words) in each language were provided for download and inspection by the participants. For some words there were multiple correct segmentations, e.g., English: pitch er s, pitcher s.

The Finnish gold standard is based on the two-level morphology analyzer FINTWOL from Lingsoft, Inc. The English gold standard is based on the CELEX English data base and the Comprehensive Grammar of the English Language by Quirk et al. (1985) The Turkish linguistic segmentations were obtained from a morphological parser developed at Bogazici University (Cetinoglu, 2000; Dutagaci, 2002). The Turkish parser is based on Oflazer's finite-state machines, with a number of changes.

## 4 Participants and their submissions

By the deadline of January 15, 2006, 12 research groups had submitted the segmentation results obtained by their algorithms. Totally 14 different algorithms were submitted and 10 of them ran experiments on all three test languages. It is noteworthy that half of the algorithms were designed by groups from the University of Leeds, where participation to this challenge was part of a course in computational linguistics. All the submitted algorithms are listed in Table 1.

In general, the submission were all interesting and relevant. Some of them failed to meet the exact specifications given, but after clarifications were requested, everyone succeeded to provide data that could be properly evaluated. The stipulated maximum count of different segments was exceeded by most of the participants, but after it turned out that this did not impede the evaluation, this restriction was removed.

In addition to the competitors' 14 segmentation algorithms, we evaluated a public baseline method called Morfessor (Creutz and Lagus, 2002; Creutz and Lagus, 2005b) organizers as well as its two more recent versions "Categories-ML" (Creutz and Lagus, 2004) and "Categories-MAP" (Creutz and Lagus, 2005a). Mikko lisaa viitteen .bibtiedostoonsa. Together with one of the challenge participants, Eric Atwell, the organizers also extended Atwell's original committee classifier algorithm "Cheat" to utilize the segmentations of all the other submissions ("Cheat-all") in addition to only the segmentations from Leeds. Naturally, these later extensions as well as the Morfessor versions competed outside the main competition and the results were included only as reference.

## 5 Competition 1

### 5.1 Evaluation

In Competition 1, for each language, the morpheme segmentations proposed by the participants' algorithm were compared against a linguistic gold standard. In the final evaluation, only a subset of all words in the data were included. For each language, a random subset consisting of 10 % of all unique word forms were picked, and the segmentations of these words were compared to the reference segmentations in the gold standard. The exact constitution of this subset was not revealed to the participants. In the evaluation, word frequency played no role. All words were equally important, were they frequent or rare.

The evaluation program, written in Perl, was provided beforehand in order to let the participants evaluate their segmentations relative to the gold standard samples provided in the Challenge. The evaluation was based on the placement of morpheme boundaries.

**Example.** Suppose that the proposed segmentation of two English words are:

boule vard
cup bearer s'

The corresponding desired (gold standard) segmentations are:

boulevard
cup bear er s '

Taken together, the proposed segmentations contain 2 hits (correctly placed boundaries between cup and bear, as well as between er and

[1] http://www.csc.fi/kielipankki/.

Table 1: The submitted algorithms.

| Name | Authors | Affiliation |
|---|---|---|
| A1 “Summaa” | Choudri and Dang | Univ. Leeds, UK |
| A2a | Bernhard | TIMC-IMAG, F |
| A2b | Bernhard | TIMC-IMAG, F |
| A3 “A.A.” | Ahmad and Allendes | Univ. Leeds, UK |
| A4a “Comb” | Bordag | Univ. Leipzig, D |
| A4b “Lsv” | Bordag | Univ. Leipzig, D |
| A5 | Rehman and Hussain | Univ. Leeds, UK |
| A6 “RePortS” | Pitler and Keshava | Univ. Yale, USA |
| A7 “Bonnier” | Bonnier | Univ. Leeds, UK |
| A8 | Kitching and Malleson | Univ. Leeds, UK |
| A9 “Pacman” | Manley and Williamson | Univ. Leeds, UK |
| A10 | Johnsen | Univ. Bergen, NO |
| A11 “Swordfish” | Jordan, Healy and Keselj | Univ. Dalhousie, CA |
| A12a “Cheat” | Atwell and Roberts | Univ. Leeds, UK |
| M1 “Baseline” | Morfessor | Helsinki Univ. Tech, FI |
| M2 “Categories-ML” | Morfessor | Helsinki Univ. Tech, FI |
| M3 “Categories-MAP” | Morfessor | Helsinki Univ. Tech, FI |
| A12b “Cheat-all” | Atwell and the organizers | Leeds and Helsinki |
| A12c “Cheat-top5” | Atwell and the organizers | Leeds and Helsinki |

`s`). There is 1 *insertion* (the incorrect boundary between `boule` and `vard`) and 2 *deletions* (the missed boundaries between `bear` and `er`, and between the plural `s` and the apostrophe `'` marking the possessive).

*Precision* is the number of hits $H$ divided by the sum of the number of hits and insertions $I$:

$$\text{Precision} = H/(H + I)\ . \tag{1}$$

*Recall* is the number of hits divided by the sum of the number of hits and deletions $D$:

$$\text{Recall} = H/(H + D)\ . \tag{2}$$

*F-Measure* is the harmonic mean of precision and recall, which equals:

$$\text{F-Measure} = 2H/(2H + I + D)\ . \tag{3}$$

According to the rules, the participant achieving the highest F-measure was to be the winner of Competition 1. In case of a tie, higher precision wins. Winners are selected separately for each language.

### 5.2 Results of Competition 1

The obtained F-measure percentages in the different tasks of Competition 1 are shown in Table 2. The corresponding precision and recall figures are shown in Tables 3 and 4, respectively.

For the Finnish task the winner (measured by F-measure) was the algorithm A2b from TIMC-IMAG in France. Next came A2a also from TIMC-IMAG and A1 from the University of Leeds. The best overall score was obtained by Morfessor M2.

A2b from TIMC-IMAG won also the Turkish task by a clear marginal. Next came A4a from the University of Leipzig and the committee classifier A12a from Leeds. The best overall score was obtained by Morfessor M3.

In the English task, the clear winner was the algorithm A6, i.e., “RePortS” from the University of Yale, who did not participate in any other language. Next came A2a and A2b from TIMC-IMAG, of which A2a scored better in this task. The A6 algorithm succeeded to beat also all Morfessors.

For English, the committee classifiers A12a, A12b, A12c from Leeds dominated all the other participants that were utilized as committee members. In Finnish only A12c and in Turkish A12a and A12c managed to do the same. Thus, the best score was always obtained by A12c, the committee of the top 5 of the other segmentation algo-

Table 2: The obtained F-measure % on different languages (Competition 1).

| Name | Finnish | Turkish | English |
|---|---|---|---|
| A1 | 61.3 | 55.4 | 49.8 |
| A2a | 63.3 | 55.3 | 66.6 |
| A2b | *64.7* | *65.3* | 62.4 |
| A3 | n.a. | n.a. | 32.0 |
| A4a | 48.3 | 57.0 | 61.7 |
| A4b | 3.8 | 5.2 | 58.5 |
| A5 | 43.4 | 45.2 | 53.8 |
| A6 | n.a. | n.a | *76.8* |
| A7 | 40.8 | 43.5 | 48.0 |
| A8 | n.a. | n.a. | 36.2 |
| A9 | 28.2 | 40.0 | 28.5 |
| A10 | n.a. | n.a. | 43.7 |
| A11 | 35.2 | 26.3 | 45.7 |
| A12a | 61.2 | 55.9 | 55.7 |
| Winner | A2b: 64.7 | A2b: 65.3 | A6: 76.8 |
| M1 | 54.2 | 51.3 | 66.0 |
| M2 | *67.0* | 69.2 | 69.0 |
| M3 | 66.4 | *70.7* | 66.2 |
| A12b | 62.0 | 59.7 | 77.4 |
| A12c | *68.3* | *71.7* | *78.6* |

Table 3: The obtained precision % on different languages (Competition 1).

| Name | Finnish | Turkish | English |
|---|---|---|---|
| A1 | 66.2 | 58.8 | 44.7 |
| A2a | 73.6 | 77.9 | 67.7 |
| A2b | 63.0 | 65.4 | 55.2 |
| A3 | n.a. | n.a. | 24.1 |
| A4a | *74.8* | *79.9* | 62.6 |
| A4b | 52.4 | 70.3 | 61.2 |
| A5 | 66.3 | 60.4 | 50.6 |
| A6 | n.a. | n.a. | *76.2* |
| A7 | 49.3 | 55.6 | 47.1 |
| A8 | n.a. | n.a. | 32.5 |
| A9 | 25.2 | 38.1 | 22.9 |
| A10 | n.a. | n.a. | 37.5 |
| A11 | 70.2 | 59.4 | 57.1 |
| A12a | 67.2 | 61.0 | 57.6 |
| Best | A4a: 74.8 | A4a: 79.9 | A6: 76.2 |
| M1 | *84.4* | 79.1 | 63.1 |
| M2 | 70.1 | 73.7 | 64.1 |
| M3 | 75.0 | 77.5 | *85.1* |
| A12b | 84.1 | *86.7* | *86.0* |
| A12c | 76.3 | 78.4 | 83.2 |

rithms.

### 5.3 Discussion

It is not that surprising that the same algorithm (A2b) wins in both the Finnish and Turkish task of Competition 1, whereas another algorithm (A6) outperforms the others in the English task. Word forming is different in Finnish and Turkish, on the one hand, and in English, on the other hand. Since English words consist of fewer morphemes, English data tends to be less sparse.

Unfortunately, the A6 algorithm, which performs extremely well on English, has not been evaluated “officially” on the two other languages. However, in their paper in these proceedings, the designers of A6 (Keshava and Pitler) report segmentation accuracies for all three languages on the small development sets provided in the challenge. It turns out that their algorithm reaches only average performance on the agglutinative languages Finnish and Turkish. Since the recall is not very high, one might assume that their algorithm suffers from the higher data sparseness of Finnish and Turkish when attempting to “peel off” prefixes and suffixes from word stems.

The committee classifier (A12a, A12b, and A12c) is an interesting approach, which generally obtains very good results. The committee classifier compares the outputs of several other systems and selects for each word the segmentation that the majority of the systems have proposed. If the majority vote results in a tie, the segmentation of the system with the highest F-measure is chosen. Thus, in order for the committee classifier to work, it seems necessary to have access to some reliable gold standard, as the performance of the other systems needs to be assessed. However, the gold standard can be fairly small, as demonstrated by the use of the segmentation samples (development sets) provided in the challenge.

## 6 Competition 2

### 6.1 Evaluation

In Competition 2, the organizers utilized the segmentations provided by the participants in order to segment the words in large corpora of Finnish as well as Turkish text. An n-gram language model was trained for this segmentation and this language model used in speech recognition experiments.

The winner of Competition 2 is the participant

Table 4: The obtained recall % on different languages (Competition 1).

| Name | Finnish | Turkish | English |
|---|---|---|---|
| A1 | 57.0 | 52.3 | 56.1 |
| A2a | 55.6 | 42.8 | 65.5 |
| A2b | *66.4* | *65.2* | 71.6 |
| A3 | n.a. | n.a. | 47.6 |
| A4a | 1.9 | 2.7 | 54.9 |
| A4b | 44.8 | 47.9 | 62.2 |
| A5 | 32.2 | 36.1 | 57.3 |
| A6 | n.a. | n.a. | *77.4* |
| A7 | 34.8 | 35.8 | 49.0 |
| A8 | n.a. | n.a. | 40.9 |
| A9 | 32.0 | 42.0 | 37.9 |
| A10 | n.a. | n.a. | 52.3 |
| A11 | 23.5 | 16.8 | 38.1 |
| A12a | 56.1 | 51.5 | 53.8 |
| Best | A2b: 66.4 | A2b: 65.2 | A6: 77.4 |
| M1 | 39.9 | 37.9 | 69.2 |
| M2 | 64.2 | 65.1 | 74.6 |
| M3 | 59.7 | 65.0 | 54.2 |
| A12b | 49.1 | 45.6 | 70.4 |
| A12c | 61.9 | *66.1* | 74.6 |

that provides the segmentation that produces the lowest letter error rate in speech recognition. The letter error is calculated as the sum of the number of substituted, inserted, and deleted letters divided by the number of letters in the correct transcription of the data.

### 6.2 Training morph-based statistical language models

The language models were trained by using exactly the same text corpus which was previously used for extracting the original word list that each competitor had processed as the competition entry. This was not a coincidence, of course, because we wanted to have segmentations for all the different word forms to be able to use the whole corpus to train the optimal sub-word language models. Naturally, we could also have tried to split any words outside the given word list using the given morph lexicon and a Viterbi search for an optimal split, as explained in (Hirsimäki et al., 2006). However, this was not needed in this case.

**Finnish.** In the Finnish newspaper, book and newswire training corpus there were totally 40 M words and 1.6 M different word forms. After splitting the whole corpus into subwords and adding the word break symbols to assist the language model, n-gram language models were trained as if the units were word sequences. The language model used resembled the traditional n-gram model as used in (Hirsimäki et al., 2006), but instead of a fixed maximum value for $n$, the $n$ was allowed to be optimized for each sequence context using the growing n-gram algorithm (Siivola and Pellom, 2005). The idea in this approach is to start from unigrams and gradually add those n-grams that maximize the training set likelihood with respect to the increase of the model size. In addition to controlling the memory consumption for training and recognition, restricting the model complexity is important also to avoid over-learning, because natural language corpora are always very sparse, even if morph units are utilized.

**Turkish.** In Turkish training corpus, there are totally 16.6 M words and 583 K different word forms. For language modeling and perplexity experiments, we used the SRI Language Modeling toolkit (Stolcke, 2002). We used interpolated modified Kneser-Ney smoothing to assign probabilities to zero probability strings. 4-gram language models are generated for each model. Entropy based pruning (Stolcke, 1998) with a pruning constant of $10^{-8}$ is applied to each model to reduce the model size.

**Model size limitation.** Despite the originally given upper limit for the lexicon size 300,000, we decided to accept submitted morph lexicons that exceeded the limit. In fact, to achieve comparable models, we only controlled the final size of the language models. For practical reasons in training and recognition, the size was set to approximately 10 million n-grams in Finnish and 50-70 Mbytes in Turkish.

### 6.3 Using cross-entropy to measure modeling accuracy

One way to directly evaluate the accuracy of a language model is to compute the average probability of an independent test text. To obtain a useful comparison measure, this probability is normalized by the number of words in the text. Typical comparison measures derived from this normalized probability are *perplexity* and *cross-entropy*. For this competition we chose cross-entropy, which is the logarithmic version (log2) of perplexity.

Given the held-out text data $T$ consisting of $W_T$

words and a language model $M$, the *cross-entropy* $H_M(T)$ was computed as:

$$H_M(T) = -\frac{1}{W_T} \log_2 P(T|M) \tag{4}$$

Here it is important that it is normalized by the number of *words*, not morphs, because a different morph lexicon was used for each model and, thus, the number of morphs in the test text varied.

Table 5: The obtained LM performance for the submitted segmentations in Finnish (Competition 2). CE is the average cross-entropy in the test text. Note that the cross-entropy is the logarithm of perplexity. As low a value as possible is desirable. OOV is the average out-of-vocabulary rate in the test text. The additional references at the bottom are explained in section 6.6.

| Finnish | CE | OOV | Lexicon size |
|---|---:|---:|---:|
| A1 | 13.65 | 0.36 | 297 981 |
| A2a | 13.54 | 0.03 | 73 178 |
| A2b | 13.63 | 0.04 | 65 557 |
| A4a | 13.55 | 2.70 | 609 458 |
| A4b | *12.93* | 0.99 | 1 559 199 |
| A5 | 13.50 | 1.24 | 650 154 |
| A7 | 13.81 | 0.85 | 530 543 |
| A9 | 13.78 | 0.95 | 615 809 |
| A11 | 13.59 | 0.58 | 690 601 |
| A12a | 13.66 | 0.40 | 317 870 |
| M1 | 13.59 | 0.02 | 121 862 |
| M2 | 13.53 | 0.08 | 155 065 |
| M3 | 13.53 | 0.16 | 164 311 |
| A12b | 13.45 | 0.47 | 355 145 |
| A12c | 13.58 | 0.14 | 171 663 |

Some additional references

| Finnish | CE | OOV | Lexicon size |
|---|---:|---:|---:|
| M1 26k | 13.62 | 0.00 | 26 935 |
| G1 | 13.62 | 0.03 | 69 929 |
| G2 | 13.31 | 0.61 | 368 412 |
| W1 | 13.95 | 0.00 | 394 266 |
| W2 | *12.04* | 5.47 | 410 001 |

Table 5 shows the obtained cross-entropies on a test text of 50,000 Finnish sentences that was randomly selected from our text corpus and held-out from the training. Although the unsupervised morph lexicons were designed to process all words, there was a small OOV (out-of-vocabulary rate) in the test text. The OOV is shown in the table, because the higher it is, the more it affects the perplexity and cross-entropy by making it look smaller than it actually would be, if the OOV was zero.

Table 6: The obtained LM performance for the submitted segmentations in Turkish (Competition 2). CE is the average cross-entropy in the test text. The OOV rate was zero, because all OOVs were split into letters. # subwords is the ratio of the number of subwords in test text to the number of words.

| Turkish | CE | # subwords | Lexicon size |
|---|---:|---:|---:|
| A1 | 15.49 | 2.92 | 121 942 |
| A2a | 14.22 | 2.42 | 48 619 |
| A2b | 15.28 | 2.87 | 37 253 |
| A4a | 14.92 | 2.66 | 204 555 |
| A4b | 14.23 | 2.23 | 561 905 |
| A5 | 15.29 | 3.03 | 195 487 |
| A7 | 14.60 | 2.61 | 189 239 |
| A9 | 16.05 | 2.89 | 218 320 |
| A11 | *13.83* | 2.04 | 264 502 |
| A12a | 15.19 | 2.77 | 148 650 |
| M1 | 13.99 | 2.30 | 51 542 |
| M2 | 14.96 | 2.79 | 96 182 |
| M3 | 14.73 | 2.70 | 88 429 |

Table 6 shows the performance on a Turkish test text consisting of 553 newspaper sentences (6989 words). If the segmentation of a test word was available in the segmentation list, we split that word into the corresponding subwords. Otherwise, the test word was left as a whole. In all of the submissions, the lexicon contained the individual letters of the Turkish alphabet as morphs. Therefore, the OOV rates were zero.

### 6.4 Large-vocabulary continuous speech recognition tests

The objective of Competition 2 was to evaluate the word splits in an application that would be as realistic as possible. When we originally planned this competition, we hesitated to choose speech recognition, because we thought it would take too much effort to build a set of state-of-art large-vocabulary continuous speech recognizers just for this evaluation. However, this was in line with our other research objectives and we have recently built several corresponding morph-based evaluation systems for Finnish, Estonian and Turkish (Hirsimäki et al., 2006; Siivola and Pellom, 2005; Kurimo et al., 2006).

**Finnish.** The speech recognizer consists of four main components: Acoustic phoneme models, language models, a lexicon and a decoder. For the acoustic models we chose the same speaker and context-dependent cross-word triphones with explicit phone duration models as for the Finnish models in (Kurimo et al., 2006) and also the same decoder (Pylkkönen, 2005). The real time factors were measured on 2.2 GHz CPU.

The lexicon and language models were created from the word splits of each competition participant and differed a little from the earlier morph models. The Finnish speech data utilized for recognizer training and evaluation was exactly the same book reading corpus as in (Hirsimäki et al., 2006; Kurimo et al., 2006). The speaker-dependent reading recognition is not the most difficult large-vocabulary recognition task as can be seen from the rather low error rates obtained, but it suits well to the scope of the Finnish language model training data and has several interesting previous benchmark results.

In a complete speech recognizer there is an almost endless amount of parameter “tweaking” in order to tune the performance, speed, memory consumption, hypothesis pruning etc., not to mention the various parameters tuned for training the models. To save effort we adopted as much as possible the same parameters as in the previous works (Hirsimäki et al., 2006; Siivola and Pellom, 2005; Kurimo et al., 2006) even if they were perhaps not exactly optimal for the new models. The only parameter that we optimized individually for each competitor was the weighting factor between the acoustic and language model to achieve the best performance on a held-out development set.

**Turkish.** The Turkish language models were evaluated by our Turkish large-vocabulary continuous speech recognizer. The main difference to the Finnish system were the speaker-independent acoustic models, the HTK frontend (Young et al., 2002) and that no explicit phone duration models were applied. The acoustic training data contained 40 hours of speech from 550 different speakers. The Turkish evaluation was performed using another decoder (Mohri and Riley, 2002) on a 2.4GHz CPU. The recognition task consisted of approximately one hour of newspaper sentences read by one female speaker.

Table 7: The obtained speech recognition performance the submitted segmentations in Finnish (Competition 2). The main measure here is the letter error rate LER. The additional references at the bottom are explained in section 6.6.

| Finnish | LER % | WER % | RTF |
|---|---|---|---|
| A1 | 1.42 | 10.58 | 17.67 |
| A2a | 1.39 | 9.53 | 12.88 |
| A2b | *1.32* | *9.47* | 15.92 |
| A4a | *1.32* | 9.81 | 15.59 |
| A4b | 1.64 | 13.54 | 10.89 |
| A5 | 1.88 | 13.10 | 13.55 |
| A7 | 1.55 | 11.33 | 13.97 |
| A9 | 1.59 | 11.71 | 16.31 |
| A11 | 1.45 | 11.17 | *10.10* |
| A12a | 1.40 | 10.72 | 15.65 |
| Winner | A2b, A4a | A2b: 9.47 | A11: 10.10 |
| M1 | 1.31 | 9.84 | 12.34 |
| M2 | 1.32 | 10.18 | 14.38 |
| M3 | *1.30* | 10.05 | 15.64 |
| A12b | 1.31 | 10.12 | 12.01 |
| A12c | *1.25* | 9.80 | 13.60 |
| Some additional references | | | |
| Finnish | LER % | WER % | RTF |
| M1 26k | 1.55 | 10.67 | 9.51 |
| G1 | 1.33 | 9.60 | 10.58 |
| G2 | 1.34 | 9.88 | 11.74 |
| W1 | 1.37 | 10.83 | 11.84 |
| W2 | 2.07 | 17.86 | *7.42* |

### 6.5 Results of Competition 2

The results of the speech recognition evaluation are shown in Table 7 (Finnish) and Table 8 (Turkish). The main performance measure is the letter error rate (LER). The word error rate (WER) was computed, too, because it is a more common measure although not so useful for the very variable-length words in Finnish. Another interesting figure is the recognition speed measured by the real-time factor (RTF).

In the Finnish task, the winners of Competition 2 were the models obtained from algorithm A2b from TIMC-IMAG in France and A4a from the University of Leipzig. The next competitors were not far behind: A2a from TIMC-IMAG, A12a and A1 from University of Leeds. The Morfessors M1, M2 and M3 were all very close to the winner. Among the top 5 models and the refer-

Table 8: The obtained speech recognition performance for the submitted segmentations in Turkish (Competition 2). The main measure here is the letter error rate LER.

| Turkish | LER % | WER % | RTF |
|---|---:|---:|---:|
| A1 | 15.0 | 43.0 | 2.68 |
| A2a | 13.6 | 38.9 | 2.15 |
| A2b | *13.4* | *37.5* | 2.19 |
| A4a | 15.7 | 46.3 | 2.43 |
| A4b | 16.7 | 50.2 | *1.75* |
| A5 | 13.5 | 38.9 | 2.46 |
| A7 | 13.8 | 40.3 | 2.33 |
| A9 | 16.9 | 47.7 | 3.03 |
| A11 | 14.6 | 41.4 | 1.85 |
| A12a | 14.5 | 41.9 | 2.56 |
| Winner | A2b: 13.4 | A2b: 37.5 | A4b: 1.75 |
| M1 | 13.7 | 39.4 | 1.98 |
| M2 | 14.3 | 41.2 | 2.10 |
| M3 | *13.2* | *37.2* | 1.89 |

ences, A2a and M1 differ from the others by being somewhat faster to run. However, the sixth best model A11 from the University of Dalhousie in Canada is clearly faster to run than the top five.

A2b from TIMC-IMAG won also Competition 2 for Turkish, but A5 from Leeds and A2a from TIMC-IMAG were very close. The Morfessor M3 produced the lowest error rates.

Since the best speech recognition error rates were not far apart, we performed the Wilcoxon's Signed-Rank test as in (Hirsimäki et al., 2006) pairwise between every algorithm pairs to see which differences are also statistically significant. For the Finnish data the best Morfessor M3 was significantly better than M1, A9, A5 and A4b. The winners of the competition A2b and A4a were both significantly better than A12a, A11, A9, A7, A5, A4b and A1.

### 6.6 Comparisons to previous references

It is also interesting to compare the current results to our earlier benchmarks. In (Hirsimäki et al., 2006) we compared pruned Morfessor baseline M1 morphs (26k and 66k lexicon) to grammatical (gold-standard) morphs (79k) and a large word-based lexicon (410k). The letter error rates in the same evaluation data were then: 4.21, 4.35, 4.57 and 6.14. However, those experiments were run in 2004 and since then we have improved the whole recognition system in many ways.

Table 9: Some additional references. In "letters" all OOVs are split to letters and in "skip" they are just left out.

| Name | Info | OOV |
|---|---|---|
| M1 26k | A small lexicon Morfessor | letters |
| G1 | Gold-standard morphs | letters |
| G2 | Gold-standard morphs | skip |
| W1 | Large word lexicon | letters |
| W2 | Large word lexicon | skip |

In (Kurimo et al., 2006) the results of the pruned Morfessor baseline M1 morphs (26k) and the large word-based lexicon (400k) in almost the same setup as in Table 7 were LER: 0.95 and 1.20; and WER: 7.0 and 8.5. The main difference was that the language models were trained such that any OOVs were modeled letter-by-letter, the training data was significantly extended (150 M words instead 40 M) and the language models were much larger (50 M n-grams instead of 10 M).

Inspired by the comparison to earlier results, we computed five additional language models for the current setup: Two from grammatical (gold-standard) morphs (79k lexicon), one pruned Morfessor baseline M1 (26k), and two large word-based lexicon (400k), see Table 9. These were all squeezed into the standard size (about 10 M n-grams) and trained with the same older (40 M) training text corpus. The results are in Table 5 and Table 7. The gold-standard morphs (G1) and the word lexicon (W1) seem to be very close in performance to the M1, but the pruned M1 (26k) has a slightly higher error rate. However, if the OOVs (the words that cannot be segmented by the lexicon) are skipped as we did for other algorithms in the Finnish part of the Competition 2, the error rates grow and cross-entropies shrink, especially for the word lexicon (W2) because of the much higher OOV rate than for any other model.

## 7 Conclusions

The objective of the Challenge was to design a statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. Ideally, these are basic vocabulary units suitable for different tasks, such as speech and text understanding, machine translation, information retrieval, and statistical language modeling.

The scientific goals of this Challenge were:

- To learn of the phenomena underlying word construction in natural languages
- To discover approaches suitable for a wide range of languages
- To advance machine learning methodology

14 different segmentation algorithms from 12 research groups were submitted and evaluated. The evaluations included 3 different languages: Finnish, Turkish and English. The algorithms and results were presented in Challenge Workshop, arranged in connection with other PASCAL Challenges on machine learning, April 10-12, 2006.

## 8 Acknowledgments

We thank all the participants for their submissions and enthusiasm. We owe great thanks as well to the organizers of the PASCAL Challenge Program who helped us organize this challenge and the challenge workshop. Our work was supported by the Academy of Finland in the projects *New information processing principles*, *Adaptive Informatics* and *New adaptive and learning methods in speech recognition*. Funding was also provided by the Graduate School of Language Technology in Finland. We thank the Finnish Federation of the Visually Impaired, Departments of Speech Science and for providing the speech data. We are also grateful to the Finnish news agency (STT) and the Finnish IT center for science (CSC) for the text data. The authors would like thank Sabanci and ODTU universities for the Turkish acoustic and text data and AT&T Labs – Research for the software. 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.

## References

M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30.

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

M. Creutz and K. Lagus. 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)*, pages 106–113, Espoo, Finland.

M. Creutz and K. Lagus. 2005b. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology. URL: http://www.cis.hut.fi/projects/morpho/.

M. Creutz and K. Linden. 2004. Morpheme segmentation gold standards for Finnish and English. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology. URL: http://www.cis.hut.fi/projects/morpho/.

Ozlem Cetinoglu 2000. Prolog based natural language processing infrastructure for Turkish. M.Sc. thesis Bogazici University, Istanbul, Turkey.

Helin Dutagaci 2002. Statistical Language Models for Large Vocabulary Continuous Speech Recognition of Turkish. M.Sc. thesis Bogazici University, Istanbul, Turkey.

P. Geutner. 1995. Using morphology towards better large-vocabulary speech recognition systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 445–448, Detroit, Michigan.

T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*. (In press).

M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, New York, USA.

Y.-S. Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, MA, USA.

M. Mohri and M. D. Riley. 2002. DCD library, speech recognition decoder library, AT&T Labs research. http://www.research.att.com/sw/tools/dcd/.

J. Pylkkönen. 2005. New pruning criteria for efficient decoding. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech)*, pages 581–584, Lisboa, Portugal.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, Essex.

V. Siivola and B. Pellom. 2005. Growing an n-gram language model. In *Proceedings of 9th European Conference on Speech Communication and Technology*.

A. Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274, Lansdowne, VA.

A. Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 901–904, Denver.

S. Young, D. Ollason, V. Valtchev, and P. Woodland. 2002. The HTK book (for HTK version 3.2.).

Y.L. Zieman and H.L. Bleich. 1997. Conceptual mapping of user's queries to medical subject headings. In *Proceedings of the 1997 American Medical Informatics Association (AMIA) Annual Fall Symposium*.