Overview and Results of Morpho Challenge 2009

Mikko Kurimo, Sami Virpioja and Ville T. Turunen
Adaptive Informatics Research Centre, Helsinki University of Technology
P.O.Box 5400, FIN-02015 TKK, Finland
Mikko.Kurimo@tkk.fi
Graeme W. Blackwood and William Byrne
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, U.K.

Abstract

In the Morpho Challenge 2009 unsupervised algorithms that provide morpheme analyses for words in different languages were evaluated in various practical applications. Morpheme analysis is particularly useful in speech recognition, information retrieval and machine translation for morphologically rich languages where the amount of different word forms is very large. The evaluations consisted of: 1. a comparison to grammatical morphemes, 2. using morphemes instead of words in information retrieval tasks, and 3. combining morpheme and word based systems in statistical machine translation tasks. The evaluation languages in 2009 were: Finnish, Turkish, German, English and Arabic. This overview paper describes the tasks, evaluation methods, and obtained results. The Morpho Challenge is part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Algorithms, Performance, Experimentation

Keywords

Morphological analysis, Machine learning

1 Introduction

Unsupervised morpheme analysis is still one of the important but unsolved tasks in computational linguistics and its applications, such as speech recognition (ASR) [3, 16], information retrieval (IR) [26, 14] and statistical machine translation (SMT) [19, 25]. The morphemes are useful, because the lexical modeling using words is particularly problematic for the morphologically rich languages, such as Finnish, Turkish and Arabic. In those languages the number of different word forms is very large because of various inflections, prefixes, suffixes and compound words.

It is possible to construct rule based tools that perform morphological analysis quite well, but of the large number of languages in the world, only few have such tools available. This is because the work of human experts to generate the rules or annotate the morpheme analysis of words and texts is expensive. Thus, learning to perform the analysis based on unannotated text collections is an important goal. Even for those languages that already have existing analysis tools, the statistical machine learning methods still propose interesting and competitive alternatives.

The scientific objectives of the Morpho Challenge are: to learn about the word construction in natural languages, to advance machine learning methodology, and to discover approaches that are suitable for many languages. In Morpho Challenge 2009, the participants first developed unsupervised algorithms and submitted their analyses for the word lists in different languages provided by the organizers. Then various evaluations were carried out using the proposed morpheme analysis to find out how they performed in different tasks. In 2009 Challenge the evaluations consisted of both a comparison to grammatical morphemes (*Competition 1*) and information retrieval and statistical machine translation tasks. The IR experiments (*Competition 2*) contained CLEF tasks, where the all the words in the queries and text corpus were replaced by their morpheme analyses. In SMT experiments (*Competition 3*) identical SMT systems using the same data are first trained using morpheme analysis and words and then combined for the best performance. The SMT tasks were first time introduced this year and are based on recent work of the organizers in morpheme based machine translation [25, 9].

2 Participants and their submissions

By the submission deadline in 8th August, 2009, ten research groups had submitted algorithms, which were then evaluated by the organizers. The authors and the names of their algorithms are listed in Table 1. The total number of tasks that the algorithms were able to participate in was 11: six in Competition 1, three in Competition 2, and two in Competition 3. The submissions for the different tasks are presented in Table 2. The final number of algorithms per task varied from 6 to 15.

Table 1: The participants and the names of their algorithms.

Author	Affiliation	Algorithm name
D. Bernhard	TU Darmstadt, D	MorphoNet
B. Can & S. Manandhar	Univ. York, UK	-
D. Currie & N. Rampersad*	Univ. Winnipeg, CA	Occam A
D. Currie & N. Rampersad*	Univ. Winnipeg, CA	Occam B
B. Golénia et al.	Univ. Bristol, UK	UNGRADE
J-F. Lavallée & P. Langlais	Univ. Montreal, CA	RALI-ANA
J-F. Lavallée & P. Langlais	Univ. Montreal, CA	RALI-COF
C. Lignos et al.	Univ. Pennsylvania & Arizona, USA	-
C. Monson et al.	Oregon Health & Science Univ., USA	ParaMor Mimic
C. Monson et al.	Oregon Health & Science Univ., USA	ParaMor-Morfessor Mimic
C. Monson et al.	Oregon Health & Science Univ., USA	ParaMor-Morfessor Union
S. Spiegler et al.	Univ. Bristol, UK	PROMODES
S. Spiegler et al.	Univ. Bristol, UK	PROMODES 2
S. Spiegler et al.	Univ. Bristol, UK	PROMODES committee
T. Tchoukalov et al.	Univ. Stanford & OHSU, USA	MetaMorph
S. Virpioja & O. Kohonen	Helsinki Univ. of Tech., FI	Allomorfessor

* The submissions were withdrawn by the request of the authors.

Statistics of the output of the submitted algorithms are briefly presented in Tables 3 – 8 for each of the languages. The average amount of analyses per word is shown in the column "#a". It is interesting that in contrary to previous years, now all algorithms ended up mostly suggesting only one analysis per word. From the column "#m" we see the average amount of morphemes per

Table 2: The submitted analyses for Arabic (non-vowelized and vowelized), English, Finnish, German and Turkish. C2 means the additional English, Finnish and German word lists for Competition 2. C3 means the Finnish and German word lists for Competition 3.

Algorithm	ara n	ara v	eng	fin	ger	tur	C2	C3
Bernhard – MorphoNet	X	X	X	X	X	X	X	X
Can & Manandhar – 1	-	-	X	-	X	X	-	-
Can & Manandhar – 2	-	-	-	-	X	X	-	-
Golénia et al. – UNGRADE	X	X	X	X	X	X	-	-
Lavallée & Langlais – RALI-ANA	X	X	X	X	X	X	-	-
Lavallée & Langlais – RALI-COF	X	X	X	X	X	X	-	-
Lignos et al.	-	-	X	-	X	-	-	-
Monson et al. – ParaMor Mimic	X	X	X	X	X	X	X	X
Monson et al. – ParaMor-Morfessor Mimic	X	X	X	X	X	X	X	X
Monson et al. – ParaMor-Morfessor Union	X	X	X	X	X	X	X	X
Spiegler et al. – PROMODES	X	X	X	X	X	X	-	-
Spiegler et al. – PROMODES 2	X	X	X	X	X	X	-	-
Spiegler et al. – PROMODES committee	X	X	X	X	X	X	-	-
Tchoukalov et al. – MetaMorph	X	X	X	X	X	X	-	X
Virpioja & Kohonen – Allomorfessor	X	X	X	X	X	X	X	X
Total	12	12	14	12	15	14	5	6

analysis, which reflects the level of details the algorithm provides. The total amount of morpheme types is given in the column "lexicon".

As baseline results for unsupervised morpheme analysis, the organizers provided morpheme analysis by a publicly available unsupervised algorithm called "Morfessor Categories-MAP" (or "Morfessor CatMAP" for short) developed at Helsinki University of Technology [6]. Analysis by the original Morfessor [5, 7] (or here "Morfessor Baseline"), which provides only a surface-level segmentation, was also provided for reference. Additionally, the reference results were provided for "letters", where the words are simply split into letters, and "Gold Standard", which is a linguistic gold standard morpheme analysis.

3 Competition 1 – Comparison to Linguistic Morphemes

3.1 Task and Data

The task was to return the given list of words in each language with the morpheme analysis added after each word. It was required that the morpheme analyses should be obtained by an unsupervised learning algorithm that would preferably be as language independent as possible. In each language, the participants were pointed to a training corpus in which all the words occur (in a sentence), so that the algorithms may also utilize information about the word context. The tasks were the same as in the Morpho Challenge 2008 last year.

The training corpora were the same as in the Morpho Challenge 2008, except for Arabic: 3 million sentences for English, Finnish and German, and 1 million sentences for Turkish in plain unannotated text files that were all downloadable from the Wortschatz collection[1] at the University of Leipzig (Germany). The corpora were specially preprocessed for the Morpho Challenge (tokenized, lower-cased, some conversion of character encodings).

For Arabic, we tried this year a very different data set, the Quran, which is smaller (only 78K words), but has also a vowelized version (as well as the unvowelized one) [22]. The corresponding full text data was also available. In Arabic, the participants could try to analyze the vowelized words or the unvowelized, or both. They were evaluated separately against the

[1] http://corpora.informatik.uni-leipzig.de/

Table 3: Statistics and example morpheme analyses in **Non-vowelized Arabic**. #a is the average amount of analyses per word (separated by a comma), #m the average amount of morphemes per analysis (separated by a space), and lexicon the total amount of morpheme types.

Algorithm	#a	#m	lexicon	example analysis
Bernhard – MorphoNet	1	2.58	4465	AdEwny A_
Golénia et al. – UNGRADE	1	3.35	7050	A d E wn y
Lavallée & Langlais – RALI-ANA	1	2.09	6192	AdEw ny
Lavallée & Langlais – RALI-COF	1	1.95	7417	AdE wA ny
Monson et al. – ParaMor Mimic	1	1.92	10255	A +dE +w +ny
Monson et al. – ParaMor-Morfessor Mimic	1	2.22	8464	A +dE +w +n y
Monson et al. – ParaMor-Morfessor Union	1	2.22	7702	AdE +w +n y
Spiegler et al. – PROMODES	1	3.66	3385	A dEwn y
Spiegler et al. – PROMODES 2	1	4.39	1136	Ad E w n y
Spiegler et al. – PROMODES committee	1	4.39	1148	Ad E w n y
Tchoukalov et al. – MetaMorph	1	1.71	10294	A dEwny
Virpioja & Kohonen – Allomorfessor	1	2.33	3546	AdE wn y
Morfessor Baseline	1	2.31	3645	AdE wn y
letters	1	5.39	37	A d E w n y
Gold Standard	1.19	10.10	10391	dEw fEl dEw +Verb +Imperative +2P +Pl +Pron +Dependent +1P

vowelized and the unvowelized gold standard analysis, respectively. For all Arabic data, the Arabic writing script were provided as well as the Roman script (Buckwalter transliteration `http://www.qamus.org/transliteration.htm`.). However, we only morpheme analysis submitted in Roman script, was evaluated.

The exact syntax of the word lists and the required output lists with the suggested morpheme analyses have been explained in [15]. As the learning is unsupervised, the returned morpheme labels may be arbitrary: e.g., "foot", "morpheme42" or "+PL". The order in which the morpheme labels appear after the word forms does not matter. Several interpretations for the same word can also be supplied, and it was left to the participants to decide whether they would be useful in the task, or not.

In Competition 1 the proposed unsupervised morpheme analyses were compared to the correct grammatical morpheme analyses called here the linguistic gold standard. The gold standard morpheme analyses were prepared in exactly the same format as the result file the participants were asked to submit, alternative analyses separated by commas. For the other languages except Arabic, the gold standard reference analyses were the same as in the Morpho Challenge 2007 [15]. For Arabic the gold standard has in each line; the word, the root, the pattern and then the morphological and part-of-speech analysis.

3.2 Evaluation

The evaluation of Competition 1 in Morpho Challenge 2009 was similar as in Morpho Challenges 2007 and 2008, but few changes were made to the evaluation measure: small bugs related to the handling of alternative analyses are fixed from the scripts, and points were now measured as one per word, not one per word pair. The data sets were the same as before for English, Finnish, German and Turkish. For Arabic, we had a new data set, the Quran, which was somewhat smaller (only 78K words) than the data set used in 2008, but has also a vowelized version (as well as the unvowelized one). The text corpus was also made available. The participants could try to analyze the vowelized words or the unvowelized, or both, and they were evaluated separately against the vowelized or the unvowelized gold standard analysis, respectively.

Because the morpheme analysis candidates are achieved by unsupervised learning, the mor-

Table 4: Vowelized Arabic statistics.

Algorithm	#a	#m	lexicon	example analysis
Bernhard – MorphoNet	1	2.84	9782	AdoEuwniy
Golénia et al. – UNGRADE	1	6.09	8523	A d o E u wni y
Lavallée & Langlais – RALI-ANA	1	2.24	11116	AdoEuw niy
Lavallée & Langlais – RALI-COF	1	1.96	12223	AdoEuwniy
Monson et al. – ParaMor Mimic	1	1.76	15875	AdoEuwniy
Monson et al. – ParaMor-Morfessor Mimic	1	2.30	13887	AdoEuwniy
Monson et al. – ParaMor-Morfessor Union	1	2.56	11096	AdoEu wniy
Spiegler et al. – PROMODES	1	5.57	4063	A d oEuwn i y
Spiegler et al. – PROMODES 2	1	7.51	726	Ad oE u w n i y
Spiegler et al. – PROMODES committee	1	6.73	1181	Ad oE uw n iy
Tchoukalov et al. – MetaMorph	1	1.99	13716	AdoEuwn iy
Virpioja & Kohonen – Allomorfessor	1	2.69	3627	AdoEuwA niy
Morfessor Baseline	1	2.89	3055	AdoEu wniy
letters	1	9.68	45	A d o E u w n i y
Gold Standard	1.19	10.45	12193	dEw faEala dEuw +Verb +Imperative +2P +Pl +Pron +Dependent +1P

pheme labels can be arbitrary and different from the ones designed by linguists. The basis of the evaluation is, thus, to compare whether any two word forms that contain the same morpheme according to the participants' algorithm also has a morpheme in common according to the gold standard and vice versa. In practice, the evaluation is performed by randomly sampling a large number of morpheme sharing word pairs from the compared analyses. Then the *precision* is calculated as the proportion of morpheme sharing word pairs in the participant's sample that really has a morpheme in common according to the gold standard. Correspondingly, the *recall* is calculated as the proportion of morpheme sharing word pairs in the gold standard sample that also exist in the participant's submission. The sample size in different languages varied depending on the size of the word lists and gold standard: 200,000 (Finnish), 50,000 (Turkish), 50,000 (German), 10,000 (English), and 5,000 (Arabic) word pairs.

Precision was calculated as follows: A number of word forms were randomly sampled from the result file provided by the participants; for each morpheme in these words, another word containing the same morpheme was chosen from the result file by random (if such a word existed). We thus obtained a number of word pairs such that in each pair at least one morpheme is shared between the words in the pair. These pairs were compared to the gold standard; a point was given if the word pair had at least the same number of common morphemes according to the gold standard as they had in the proposed analysis. If the gold standard had common morphemes, but less than proposed, fractions of points were given. In the case of alternative analyses in the gold standard, the best matching alternative was used. The maximum number of points for one sampled word was normalized to one. The total number of points was then divided by the total number of sampled words.

For instance, assume that the proposed analysis of the English word "abyss" is "abys +s". Two word pairs are formed: Say that "abyss" happens to share the morpheme "abys" with the word "abysses"; we thus obtain the word pair "abyss - abysses". Also assume that "abyss" shares the morpheme "+s" with the word "mountains"; this produces the pair "abyss - mountains". Now, according to the gold standard the correct analyses of these words are: "abyss_N", "abyss_N +PL", "mountain_N +PL", respectively. The pair "abyss - abysses" is correct (common morpheme: "abyss_N"), but the pair "abyss - mountain" is incorrect (no morpheme in common). Precision for the word "abyss" is thus $1/2 = 50\%$.

For words that had several alternative analyses, as well as for word pairs that have more than one morpheme in common, normalization of the points was carried out. In short, an equal weight

Table 5: English statistics.

Algorithm	#a	#m	lexicon	example analysis
Bernhard – MorphoNet	1	1.75	211439	vulnerabilty _ies
Can & Manandhar	1	2.09	150097	vulner abilities
Golénia et al. – UNGRADE	1	3.87	123634	vulnerabilities
Lavallée & Langlais – RALI-ANA	1	2.10	166826	vulner abiliti es
Lavallée & Langlais – RALI-COF	1	1.91	145733	vulnerability ies
Lignos et al.	1	1.74	198546	VULNERABILITY +(ies)
Monson et al. – ParaMor Mimic	1	3.03	188716	vulner +a +bilit +ie +s
Monson et al. – ParaMor-Morfessor Mimic	1	2.96	166310	vulner +a +bilit +ies
Monson et al. – ParaMor-Morfessor Union	1	2.87	120148	vulner a +bilit +ies
Spiegler et al. – PROMODES	1	3.28	107111	vul nerabilitie s
Spiegler et al. – PROMODES 2	1	3.63	47456	v ul nera b ili ties
Spiegler et al. – PROMODES committee	1	3.63	47456	v ul nera b ili ties
Tchoukalov et al. – MetaMorph	1	1.58	241013	vulnerabiliti es
Virpioja & Kohonen – Allomorfessor	1	2.59	23741	vulnerability ies
Morfessor Baseline	1	2.31	40293	vulner abilities
Morfessor CatMAP	1	2.12	132038	vulner abilities
letters	1	9.10	28	v u l n e r a b i l i t i e s
Gold Standard	1.06	2.49	18855	vulnerable_A ity_s +PL

is given for each alternative analysis, as well as each word pair in an analysis. E.g., if a word has three alternative analyses, the first analysis has four morphemes, and the first word pair in that analysis has two morphemes in common, each of the two common morphemes will amount to $1/3 * 1/4 * 1/2 = 1/24$ of the one point available for that word.

Recall was calculated analogously to precision: A number of word forms were randomly sampled from the gold standard file; for each morpheme in these words, another word containing the same morpheme was chosen from the gold standard by random (if such a word existed). The word pairs were then compared to the analyses provided by the participants; a full point was given for each sampled word pair that had at least as many morphemes in common also in the analyses proposed by the participants' algorithm. Again, points per word was normalized to one and the total number of points was divided by the total number of words.

The *F-measure*, which is the harmonic mean of precision and recall, was selected as the final evaluation measure:

$$\text{F-measure} = 1/(1/\text{Precision} + 1/\text{Recall})\,. \tag{1}$$

3.3 Results

The results of the Competition 1 are presented in Tables 9–14. In three languages, Turkish, Finnish and German, the algorithms with the clearly highest F-measure were "ParaMor-Morfessor Mimic" and "Union". In English, however, "Allomorfessor" was better and also the algorithm by Lignos et al. was quite close. In Arabic, the results turned out quite surprising, because most algorithms gave rather low recall and F-measure and nobody was able to beat the simple "letters" reference. "Promodes" and "Ungrade" methods scored clearly better than the rest of the participants in Arabic.

The tables contain also results of the best algorithms from Morpho Challenges 2008 [18] and 2007 [15]. From Morpho Challenge 2008, the best method "Paramor + Morfessor" would have also scored highest in 2009. However, "Paramor + Morfessor" was a combination of two separate algorithms, ParaMor and Morfessor, where the two different analyses were just given as alternative analyses for each word. As the evaluation procedure selects the best matching analysis, this boosts up the recall, while obtaining precision that is about the average of the two algorithms. By combining this year's top algorithms in a similar manner, it would be easy to get even higher scores. However, exploiting this property of the evaluation measure is not a very interesting approach.

Table 6: Finnish statistics.

Algorithm	#a	#m	lexicon	example analysis
Bernhard – MorphoNet	1	2.53	984581	eu-jäsenmaita _ss_ _-_ _s_ _ssa
Golénia et al. – UNGRADE	1	4.60	790814	eu jäse nmaiss a
Lavallée & Langlais – RALI-ANA	1	2.05	1217550	eu- jäsenmais s a
Lavallée & Langlais – RALI-COF	1	2.39	723171	jäsenmaissa eu-
Monson et al. – ParaMor Mimic	1	3.30	1149382	eu-jäsenma +i +ssa
Monson et al. – ParaMor-Morfessor Mimic	1	4.24	323561	eu- +jäsen ma +i +ssa
Monson et al. – ParaMor-Morfessor Union	1	4.02	215118	eu- +jäsen ma +i +ssa
Spiegler et al. – PROMODES	1	4.88	296010	e u jä sen mais sa
Spiegler et al. – PROMODES 2	1	5.64	97558	eu j äsen mais sa
Spiegler et al. – PROMODES committee	1	4.94	199700	e u j äsen maissa
Tchoukalov et al. – MetaMorph	1	1.87	2036496	eu- jäsenmaissa
Virpioja & Kohonen – Allomorfessor	1	2.46	70228	eu- jäsen maissa
Morfessor Baseline	1	2.21	149417	eu- jäsenmaissa
Morfessor CatMAP	1	2.94	217001	eu- jäsen maissa
letters	1	13.78	32	e u - j ä s e n m a i s s a
Gold Standard	1.16	3.52	41815	eu jäsen_N maa_N +PL +INE

Excluding "Paramor + Morfessor", this year's best scores for the English, Finnish, German and Turkish tasks are higher than the best scores in 2008. However, Bernhard's second method from 2007 holds still the highest score for English, Finnish and German. The best result for the Turkish task has improved yearly.

4 Competition 2 – Information Retrieval

In Competition 2, the morpheme analyses were compared by using them in an Information Retrieval (IR) task with three languages: English, German and Finnish. The Competition 2 IR tasks and corpora were the same as in our previous Morpho Challenges in 2007 [14] and 2008 [17]. The participants were asked to submit segmentation for the given word lists. In the evaluation, words occurring in the corpus and the queries were replaced by the morpheme segmentations in the submitted word lists. Additionally, there was an option to access the test corpus and evaluate the IR performance using the morpheme analysis of word forms in their full text context.

Morpheme analysis is important in a text retrieval task because the user will want to retrieve all documents irrespective of which word forms are used in the query and in the text. Of the tested languages, Finnish is the most complex morphologically and is expected to gain most from a successful analysis. Compound words are typical of German while English is morphologically the simplest.

The participants' submissions were compared against a number of reference methods. Like the participants' methods, Morfessor baseline [4, 7] and Morfessor Categories-MAP [6] are unsupervised algorithms. Also evaluated were a commercial word normalization tool (TWOL) and the rule-based grammatical morpheme analyses based on the linguistic gold standards [8]. These methods have the benefit of language specific linguistic knowledge embedded in them. Traditional stemming approaches based on the Porter stemmer [21] as well as using the words without any processing were also tested.

4.1 Task and Data

In a text retrieval task, the user formulates their information need to a query and the system has to return all documents from the collection that satisfy the user's infomation need. To evaluate

Table 7: German statistics.

Algorithm	#a	#m	lexicon	example analysis
Bernhard – MorphoNet	1	1.57	816818	durchlief _en
Can & Manandhar - 1	1	2.27	320971	durch liefen
Can & Manandhar - 2	1	2.56	274334	durch lief en
Golénia et al. – UNGRADE	1	4.17	497947	durc hliefe n
Lavallée & Langlais – RALI-ANA	1	1.93	695995	durchlief en
Lavallée & Langlais – RALI-COF	1	2.17	429399	liefe durch n
Lignos et al.	1	2.08	515357	DURCHLIEF +(en)
Monson et al. – ParaMor Mimic	1	2.74	726045	durchlief +en
Monson et al. – ParaMor-Morfessor Mimic	1	3.87	206380	durch lief +en
Monson et al. – ParaMor-Morfessor Union	1	3.63	156981	durch lief +en
Spiegler et al. – PROMODES	1	3.67	326224	durchliefen
Spiegler et al. – PROMODES 2	1	4.97	94057	dur chlie fen
Spiegler et al. – PROMODES committee	1	4.13	183007	durchlie fen
Tchoukalov et al. – MetaMorph	1	1.96	848660	du rchliefen
Virpioja & Kohonen – Allomorfessor	1	2.63	43609	durch liefen
Morfessor Baseline	1	2.30	90009	durch liefen
Morfessor CatMAP	1	3.06	172907	durch liefen
letters	1	13.39	29	d u r c h l i e f e n
Gold Standard	1.19	3.32	16215	durch_P lauf_V +PAST +13PL

the performance of a retrieval system, a collection of documents, a number of test queries and a set of human relevance assessments are needed.

In Competition 2, the participants' only task was to provide segmentations for the given word lists. The word lists were extracted from the test corpora and queries. In addition, the words in the Competition 1 word lists were added to the Competition 2 lists. Optionally, the participants could also register to the Cross-Language Evaluation Forum (CLEF)[2] and use the full text corpora for preparing the morpheme analysis. The IR experiments were performed by the Morpho Challenge organizers by using the submitted word lists to replace the words both in the documents and in the queries by their proposed analyses.

The corpora, queries and relevance assessments were provided by CLEF and contained news paper articles as follows:

- In Finnish: 55K documents from short articles in Aamulehti 1994-95, 50 test queries on specific news topics and 23K binary relevance assessments (CLEF 2004)
- In English: 170K documents from short articles in Los Angeles Times 1994 and Glasgow Herald 1995, 50 test queries on specific news topics and 20K binary relevance assessments (CLEF 2005).
- In German: 300K documents from short articles in Frankfurter Rundschau 1994, Der Spiegel 1994-95 and SDA German 1994-95, 60 test queries with 23K binary relevance assessments (CLEF 2003).

4.2 Reference methods

The performance of the participating algorithms was compared to a number of reference methods. Some these methods are commonly used in IR and the purpose of providing these methods is to evaluate the usefulness of the unsupervised algorithms for the task. The reference methods are the same as used in Morpho Challenge 2008 [17].

[2] http://www.clef-campaign.org/

Table 8: Turkish statistics.

Algorithm	#a	#m	lexicon	example analysis
Bernhard – MorphoNet	1	3.45	180103	Cukur _I_ _yl_ _Iyl_ _yla _Iyla
Can & Manandhar - 1	1	1.94	326178	CukurlarIyla
Can & Manandhar - 2	1	3.84	208317	Cu kurlarI y la
Golénia et al. – UNGRADE	1	3.75	211236	C u kur la rIyl a
Lavallée & Langlais – RALI-ANA	1	2.35	193643	Cuk urlarIyla
Lavallée & Langlais – RALI-COF	1	3.68	83624	tur Cuk lar I yla
Monson et al. – ParaMor Mimic	1	3.69	218334	Cukur +lar +Iyla
Monson et al. – ParaMor-Morfessor Mimic	1	4.05	150775	Cukur +lar +I +yla
Monson et al. – ParaMor-Morfessor Union	1	3.86	104868	Cukur +lar +I +yla
Spiegler et al. – PROMODES	1	5.16	62869	C ukurlarIyl a
Spiegler et al. – PROMODES 2	1	5.12	14364	C u kur larIy la
Spiegler et al. – PROMODES committee	1	3.38	113584	C u kurlarI y la
Tchoukalov et al. – MetaMorph	1	1.99	470080	Cu kurlarIyla
Virpioja & Kohonen – Allomorfessor	1	2.37	29193	Cukur larIyla
Morfessor Baseline	1	2.14	53473	Cukur larIyla
Morfessor CatMAP	1	2.64	114834	Cukur +larI +yla
letters	1	9.99	33	C u k u r l a r I y l a
Gold Standard	1.96	3.53	26151	Cukur +PL +POS3 +REL, Cukur +POS3S +REL

1. *Morfessor Categories-MAP*: The Morfessor Categories-MAP (or here just "CatMAP", for short) was used for the unsupervised morpheme analysis. The stem vs. suffix tags were kept, but did not receive any special treatment in the indexing as we wanted to keep the IR evaluation as unsupervised as possible.

2. *Morfessor Baseline*: Morfessor Baseline algorithm was used to split words into smaller pieces without any real morpheme analysis. This means that all the obtained subword units were directly used as index terms.

3. *dummy*: No segmentation or analysis was performed and words were used as index terms as such. The only processing step was that hyphens were replaced by spaces so that hyphenated words were indexed as separate words. We expected that although the morpheme analysis should provide helpful information for IR, all the submissions would not probably be able to beat this simple baseline. However, if some morpheme analysis method would consistently beat this baseline in all languages and task, it would mean that the method would probably be useful in a language and task independent way.

4. *grammatical*: The words were analyzed using the same gold standard analyses in each language that were utilized as the "ground truth" in the Competition 1. Besides the stems and suffixes, the gold standard analyses typically consist of all kinds of grammatical tags which we decided to simply include as index terms, as well. For many words the gold standard analyses included several alternative interpretations. We tried two approaches to deal with that fact. Either only the first interpretation was used ("grammatical first") or all of them ("grammatical all"). Words that were not in the gold standard segmentation were indexed as such. Because our gold standards are quite small, 60k (English) - 600k (Finnish), compared to the amount of words that the unsupervised methods can analyze, we did not expect "grammatical" to perform particularly well, even though it would probably capture some useful indexing features to beat the "dummy", at least.

5. *snowball*: No real morpheme analysis was performed, but the words were stemmed by language specific stemming algorithms provided by Snowball libstemmer library. Porter stemming algorithm was used for English. Finnish and German stemmers were used for the other

Table 9: The submitted unsupervised morpheme analyses compared to the gold standard in **non-vowelized Arabic** (Competition 1).

Author	Method	Precision	Recall	F-measure
-	letters	70.48%	53.51%	60.83%
Spiegler et al.	PROMODES 2	76.96%	37.02%	50.00%
Spiegler et al.	PROMODES committee	77.06%	36.96%	49.96%
Spiegler et al.	PROMODES	81.10%	20.57%	32.82%
Golénia et al.	UNGRADE	83.48%	15.95%	26.78%
Virpioja & Kohonen	Allomorfessor	91.62%	6.59%	12.30%
-	Morfessor Baseline	91.77%	6.44%	12.03%
Bernhard	MorphoNet	90.49%	4.95%	9.39%
Monson et al.	ParaMor-Morfessor Union	93.72%	4.81%	9.14%
Monson et al.	ParaMor-Morfessor Mimic	93.76%	4.55%	8.67%
Lavallée & Langlais	RALI-ANA	92.40%	4.40%	8.41%
Tchoukalov et al.	MetaMorph	95.05%	2.72%	5.29%
Monson et al.	ParaMor Mimic	91.29%	2.56%	4.97%
Lavallée & Langlais	RALI-COF	94.56%	2.13%	4.18%

Table 10: The submitted unsupervised morpheme analyses compared to the gold standard in **vowelized Arabic** (Competition 1).

Author	Method	Precision	Recall	F-measure
-	letters	50.56%	84.08%	63.15%
Spiegler et al.	PROMODES 2	63.00%	59.07%	60.97%
Spiegler et al.	PROMODES committee	68.32%	47.97%	56.36%
Golénia et al.	UNGRADE	72.15%	43.61%	54.36%
Spiegler et al.	PROMODES	74.85%	35.00%	47.70%
-	Morfessor Baseline	86.87%	4.90%	9.28%
Monson et al.	ParaMor-Morfessor Union	91.61%	4.41%	8.42%
Virpioja & Kohonen	Allomorfessor	88.28%	4.37%	8.33%
Monson et al.	ParaMor-Morfessor Mimic	93.62%	3.23%	6.24%
Bernhard	MorphoNet	92.52%	2.91%	5.65%
Tchoukalov et al.	MetaMorph	88.78%	2.89%	5.59%
Lavallée & Langlais	RALI-ANA	91.30%	2.83%	5.49%
Monson et al.	ParaMor Mimic	91.36%	1.85%	3.63%
Lavallée & Langlais	RALI-COF	95.09%	1.50%	2.95%

Table 11: The submitted unsupervised morpheme analyses compared to the gold standard in **English** (Competition 1).

Author	Method	Precision	Recall	F-measure
Virpioja & Kohonen	Allomorfessor	68.98%	56.82%	62.31%
-	Morfessor Baseline	74.93%	49.81%	59.84%
Monson et al.	ParaMor-Morfessor Union	55.68%	62.33%	58.82%
Lignos et al.	-	83.49%	45.00%	58.48%
Monson et al.	ParaMor Mimic	53.13%	59.01%	55.91%
Bernhard	MorphoNet	65.08%	47.82%	55.13%
Monson et al.	ParaMor-Morfessor Mimic	54.80%	60.17%	57.36%
Lavallée & Langlais	RALI-COF	68.32%	46.45%	55.30%
Can & Manandhar	-	58.52%	44.82%	50.76%
-	Morfessor CatMAP	84.75%	35.97%	50.50%
Spiegler et al	PROMODES	36.20%	64.81%	46.46%
Lavallée & Langlais	RALI-ANA	64.61%	33.48%	44.10%
Spiegler et al.	PROMODES 2	32.24%	61.10%	42.21%
Spiegler et al.	PROMODES committee	32.24%	61.10%	42.21%
Tchoukalov et al.	MetaMorph	68.41%	27.55%	39.29%
Golénia et al.	UNGRADE	28.29%	51.74%	36.58%
-	letters	3.82%	99.88%	7.35%
Monson et al. 2008	ParaMor + Morfessor	69.59%	65.57%	67.52%
Monson et al. 2008	ParaMor	63.32%	51.96%	57.08%
Bernhard 2007	2	67.42%	65.11%	66.24%

Table 12: The submitted unsupervised morpheme analyses compared to the gold standard in **Finnish** (Competition 1).

Author	Method	Precision	Recall	F-measure
Monson et al.	ParaMor-Morfessor Union	47.89%	50.98%	49.39%
Monson et al.	ParaMor-Morfessor Mimic	51.75%	45.42%	48.38%
-	Morfessor CatMAP	79.01%	31.08%	44.61%
Spiegler et al.	PROMODES committee	41.20%	48.22%	44.44%
Monson et al.	ParaMor Mimic	47.15%	40.50%	43.57%
Spiegler et al.	PROMODES 2	33.51%	61.32%	43.34%
Spiegler et al.	PROMODES	35.86%	51.41%	42.25%
Lavallée & Langlais	RALI-COF	74.76%	26.20%	38.81%
Golénia et al.	UNGRADE	40.78%	33.02%	36.49%
Bernhard	MorphoNet	63.35%	22.62%	33.34%
Virpioja & Kohonen	Allomorfessor	86.51%	19.96%	32.44%
-	Morfessor Baseline	89.41%	15.73%	26.75%
Tchoukalov et al.	MetaMorph	37.17%	15.15%	21.53%
Lavallée & Langlais	RALI-ANA	60.06%	10.33%	17.63%
-	letters	5.17%	99.89%	9.83%
Monson et al. 2008	ParaMor + Morfessor	65.21%	50.43%	56.87%
Monson et al. 2008	ParaMor	49.97%	37.64%	42.93%
Bernhard 2007	2	63.92%	44.48%	52.45%

Table 13: The submitted unsupervised morpheme analyses compared to the gold standard in **German** (Competition 1).

Author	Method	Precision	Recall	F-measure
Monson et al.	ParaMor-Morfessor Union	52.53%	60.27%	56.14%
Monson et al.	ParaMor-Morfessor Mimic	51.07%	57.79%	54.22%
-	Morfessor CatMAP	71.08%	38.92%	50.30%
Monson et al.	ParaMor Mimic	50.81%	47.68%	49.20%
Can & Manandhar	2	57.67%	42.67%	49.05%
Lavallée & Langlais	RALI-COF	67.53%	34.38%	45.57%
Spiegler et al.	PROMODES 2	36.11%	50.52%	42.12%
Virpioja & Kohonen	Allomorfessor	77.78%	28.83%	42.07%
Bernhard	MorphoNet	67.41%	30.19%	41.71%
Spiegler et al.	PROMODES	49.88%	33.95%	40.40%
Spiegler et al.	PROMODES committee	48.48%	34.61%	40.39%
-	Morfessor Baseline	81.70%	22.98%	35.87%
Lignos et al.	-	78.90%	21.35%	33.61%
Golénia et al.	UNGRADE	39.02%	29.25%	33.44%
Tchoukalov et al.	MetaMorph	39.59%	19.81%	26.40%
Can & Manandhar	1	73.16%	15.27%	25.27%
Lavallée & Langlais	RALI-ANA	61.39%	15.34%	24.55%
-	letters	2.79%	99.92%	5.43%
Monson et al. 2008	ParaMor + Morfessor	64.06%	61.52%	62.76%
Monson et al. 2008	ParaMor	70.73%	38.82%	50.13%
Bernhard 2007	2	54.02%	60.77%	57.20%

Table 14: The submitted unsupervised morpheme analyses compared to the gold standard in **Turkish** (Competition 1).

Author	Method	Precision	Recall	F-measure
Monson et al.	ParaMor-Morfessor Mimic	48.07%	60.39%	53.53%
Monson et al.	ParaMor-Morfessor Union	47.25%	60.01%	52.88%
Monson et al.	ParaMorMimic	49.54%	54.77%	52.02%
Lavallée & Langlais	RALI-COF	48.43%	44.54%	46.40%
-	Morfessor CatMAP	79.38%	31.88%	45.49%
Spiegler et al.	PROMODES 2	35.36%	58.70%	44.14%
Spiegler et al.	PROMODES	32.22%	66.42%	43.39%
Bernhard	MorphoNet	61.75%	30.90%	41.19%
Can & Manandhar	2	41.39%	38.13%	39.70%
Spiegler et al.	PROMODES committee	55.30%	28.35%	37.48%
Golénia et al.	UNGRADE	46.67%	30.16%	36.64%
Tchoukalov et al.	MetaMorph	39.14%	29.45%	33.61%
Virpioja & Kohonen	Allomorfessor	85.89%	19.53%	31.82%
-	Morfessor Baseline	89.68%	17.78%	29.67%
Lavallée & Langlais	RALI-ANA	69.52%	12.85%	21.69%
-	letters	8.66%	99.13%	15.93%
Can & Manandhar	1	73.03%	8.89%	15.86%
Monson et al. 2008	ParaMor + Morfessor	66.78%	57.97%	62.07%
Monson et al. 2008	ParaMor	57.35%	45.75%	50.90%
Bordag 2007	5a	81.06%	23.51%	36.45%

languages. Hyphenated words were first split to parts that were then stemmed separately. Stemming is expected to perform very well for English but not necessarily for the other languages because for them it is harder to find good stems.

6. *TWOL*: Two-level morphological analyzer TWOL from Lingsoft Inc.[3] was used to find the normalized forms of the words which were then used as index terms. Some words may have several alternative interpretations and two cases were studied similarly to the grammatical case. Either all alternatives were used ("TWOL all") or only the first one ("TWOL first"). Compound words were split to parts. Words not recognized by the analyzer were indexed as such. This method is expected to perform very well because of the language specific linguistic knowledge used.

7. *Best2008*: This is the algorithm in each task that provided the highest average precision in Morpho Challenge 2008. The IR tasks in 2009 were identical to 2008.

4.3 Evaluation

English, German and Finnish IR tasks were used to evaluate the submitted morpheme analyses. Unfortunately, neither Turkish or Arabic IR test corpora were available for the organizers. The experiments were performed by replacing the words in the corpus and the queries by the submitted morpheme analyses. Thus, the retrieval was based on morphemes as index terms. If a segmentation for a word was not provided, it was left unsegmented and used as a separate morpheme. The queries were formed by using the title and description ("TD") fields from the topic descriptions.

The IR experiments were performed using the freely available LEMUR toolkit[4] version 4.4. The popular Okapi BM25 ranking function was used. In the 2007 challenge [14], it was noted that the performance of Okapi BM25 suffers greatly if the corpus contains morphemes that are very common. The unsupervised morpheme segmentation algorithms tend to introduce such morphemes when they e.g. separate suffixes. To overcome this problem, a method for automatically generating a stoplist was introduced. Any term that has a collection frequency higher than 75000 (Finnish) or 150000 (German and English) is added to the stoplist and thus excluded from the corpus. Even though the method is quite simplistic, it generates reasonable sized stoplists (about 50-200 terms) and is robust with respect to the cutoff parameter. With a stoplist, Okapi BM25 clearly outperformed TFIDF ranking and thus the approach has been adopted for later evaluations as well. The evaluation criterion for the IR performance is the Mean Average Precision (MAP) that was calculated using the `trec_eval` program.

4.4 Results

Three research groups submitted total of five different segmentations for the Competition 2 word lists. In addition, for the 6 groups and 10 algorithms that did not provide segmentations for the Competition 2 word lists, the smaller Competition 1 word list was used. None of the participants used the option to use the full text corpora to provide analyses for words in their context.

Tables 15, 16 and 17 show the obtained MAP values for the submissions in English, German and Finnish respectively. For English, the best performance was achieved by the algorithm by Lignos et al. even though only the shorter Competition 1 word list was available for evaluation. "ParaMor-Morfessor Mimic" and "ParaMor-Morfessor Union" by Monson et. al gave the best performance for German and Finnish respectively. Overall, the algorithms by Monson et al., especially "ParaMor-Morfessor Union", gave good performance across all tested languages. Also, "Allomorfessor" by Virpioja & Kohonen was a solid performer in all languages. However, none of the submitted algorithms could beat the winners of last year's competition.

In all languages, the best performance was achieved by one of the reference algorithms. The rule based word normalizer, TWOL, gave best performance in German and Finnish. In the

[3] `http://www.lingsoft.fi/`

[4] `http://www.lemurproject.org/`

Table 15: The obtained mean average precision (MAP) in the information retrieval task for **English**. Asterisk (*) denotes submissions that did not include segmentations for Competition 2 and were evaluated by using the shorter Competition 1 word list.

Author	Method	MAP
-	snowball porter	0.4081
-	Best2008 (Monson Paramor+Morfessor)	0.3989
-	TWOL first	0.3957
-	TWOL all	0.3922
Lignos et al.	-	0.3890*
-	Morfessor Baseline	0.3861
Virpioja & Kohonen	Allomorfessor	0.3852
Monson et al.	ParaMor Mimic	0.3822
Monson et al.	ParaMor-Morfessor Union	0.3811
-	grammatical first	0.3734
-	Morfessor CatMAP	0.3713
Lavellée & Langlais	RALI-ANA	0.3707*
Monson et al.	ParaMor-Morfessor Mimic	0.3649
Tchoukalov et al.	MetaMorph	0.3623*
Lavellée & Langlais	RALI-COF	0.3616*
Bernhard	MorphoNet	0.3560
-	grammatical all	0.3542
-	dummy	0.3293
Golénia et al.	UNGRADE	0.2996*
Can & Manandhar	-	0.2940*
Spiegler et al.	PROMODES	0.2917*
Spiegler et al.	PROMODES 2	0.2066*
Spiegler et al.	PROMODES committee	0.2066*

Table 16: The obtained mean average precision (MAP) in the information retrieval task for **German**. Asterisk (*) denotes submissions that did not include segmentations for Competition 2 and were evaluated by using the shorter Competition 1 word list.

Author	Method	MAP
-	TWOL first	0.4885
-	TWOL all	0.4743
-	Best2008 (Monson Paramor+Morfessor)	0.4734
-	Morfessor Baseline	0.4656
-	Morfessor CatMAP	0.4642
Monson et al.	ParaMor-Morfessor Mimic	0.4490
Monson et al.	ParaMor-Morfessor Union	0.4478
Virpioja & Kohonen	Allomorfessor	0.4388
Can & Manandhar	1	0.4006*
Lavellée & Langlais	RALI-COF	0.3965*
Can & Manandhar	2	0.3952*
-	snowball german	0.3865
Lignos et al.	-	0.3863*
Monson et al.	ParaMor Mimic	0.3757
Tchoukalov et al.	MetaMorph	0.3752*
Spiegler et al.	PROMODES committee	0.3634*
-	dummy	0.3509
Golénia et al.	UNGRADE	0.3496*
Spiegler et al.	PROMODES	0.3484*
-	grammatical first	0.3353
Lavellée & Langlais	RALI-ANA	0.3284*
Bernhard	MorphoNet	0.3167
-	grammatical all	0.3014
Spiegler et al.	PROMODES 2	0.2997*

Table 17: The obtained mean average precision (MAP) in the information retrieval task for **Finnish**. Asterisk (*) denotes submissions that did not include segmentations for Competition 2 and were evaluated by using the shorter Competition 1 word list.

Author	Method	MAP
-	TWOL first	0.4976
-	Best2008 (McNamee four)	0.4918
-	TWOL all	0.4845
Monson et al.	ParaMor-Morfessor Union	0.4713
Virpioja & Kohonen	Allomorfessor	0.4601
-	Morfessor CatMAP	0.4441
-	Morfessor Baseline	0.4425
-	grammatical first	0.4312
Monson et al.	ParaMor-Morfessor Mimic	0.4294
-	snowball finnish	0.4275
-	grammatical all	0.4090
Spiegler et al.	PROMODES 2	0.3857*
Monson et al.	ParaMor Mimic	0.3819
Lavellée & Langlais	RALI-COF	0.3740*
Bernhard	MorphoNet	0.3668
Golénia et al.	UNGRADE	0.3636*
Lavellée & Langlais	RALI-ANA	0.3595*
-	dummy	0.3519
Spiegler et al.	PROMODES committee	0.3492*
Spiegler et al.	PROMODES	0.3392*
Tchoukalov et al.	MetaMorph	0.3289*

English task, TWOL was only narrowly beaten by the traditional Porter stemmer. For German and Finnish, stemming was not nearly as efficient. Of the other reference methods, "Morfessor Baseline" gave good performance in all languages while the "grammatical" reference based on linguistic analyses did not perform well probably because the gold standards are quite small.

4.5 Statistical testing

For practical reasons, a limited set of queries (50-60) are used in evaluation of the IR-performance. The obtained results will include variation between queries as well as between methods. Statistical testing was employed to determine what differences in performance between the submissions are greater than expected by pure chance. The methodology we use follows closely the one used in TREC [10] and CLEF [1].

Analysis was performed with Two-way ANOVA using MATLAB Statistics Toolbox. Since ANOVA assumes the samples to be normally distributed, a transformation for the average precision values was made with the arcsin-root function:

$$f(x) = \arcsin(\sqrt{x}). \tag{2}$$

The transformation makes the samples more normally distributed. Statistical significances were examined using MATLAB's `multcompare` function with the Tukey t-test and 0.05 confidence level.

Results of the test are summarized for English, German and Finnish in Figures 1, 2 and 3 respectively. Participant or reference submission name is shown on the y-axis and the performance on the x-axis. The average performance of the method is indicated by a circle and the bars show the confidence interval in which the difference in performance is not statistically significant. The "top group" or the submissions that have no significant difference to the best result of each language is highlighted.

The confidence intervals are relatively wide and a large proportion of the submissions are in the top group for all languages. It is well known and also noted in the CLEF Ad Hoc track [1] that it is hard to obtain statistically significant differences between retrieval results with only 50 queries.

One interesting comparison is to see if there are significant differences to the "dummy" case where no morphological analysis is performed. For German and Finnish, "ParaMor-Morfessor Union" is the only submission that is significantly better than the dummy method. For English, none of the participants' results can significantly improve over "dummy". Only the Porter stemmer is significantly better according to the test.

4.6 Discussions

The results of the Competition 2 suggest that unsupervised morphological analysis is a viable approach for information retrieval. Some of the unsupervised methods were able to beat the "dummy" baseline and the best were close to the language specific rule-based "TWOL" word normalizer. However, this year's competition did not offer any improvements to previous results.

The fact that segmentations of the full Competition 2 word list was not provided by all participants makes the comparison of IR performance a bit more difficult. The participants that were evaluated using only the Competition 1 word lists had a disadvantage, because then the additional words in the IR task were indexed as such without analysis. In the experiments in Morpho Challenge 2007 [14], the segmentation of the additional words improved performance in the Finnish task for almost all participants. In German and English tasks the improvements were small. However, if the segmentation algorithm is not performing well, leaving some of the words unsegmented only improves the results for that participant.

Most of the methods that performed well in the Competition 2 IR task were also strong in the corresponding linguistic evaluation of Competition 1 and vice versa. The biggest exeptions were in the Finnish task where the "PROMODES committee" algorithm gave reasonably good results in the linguistic evaluation but not in the IR task. The algorithm seems to oversegment words

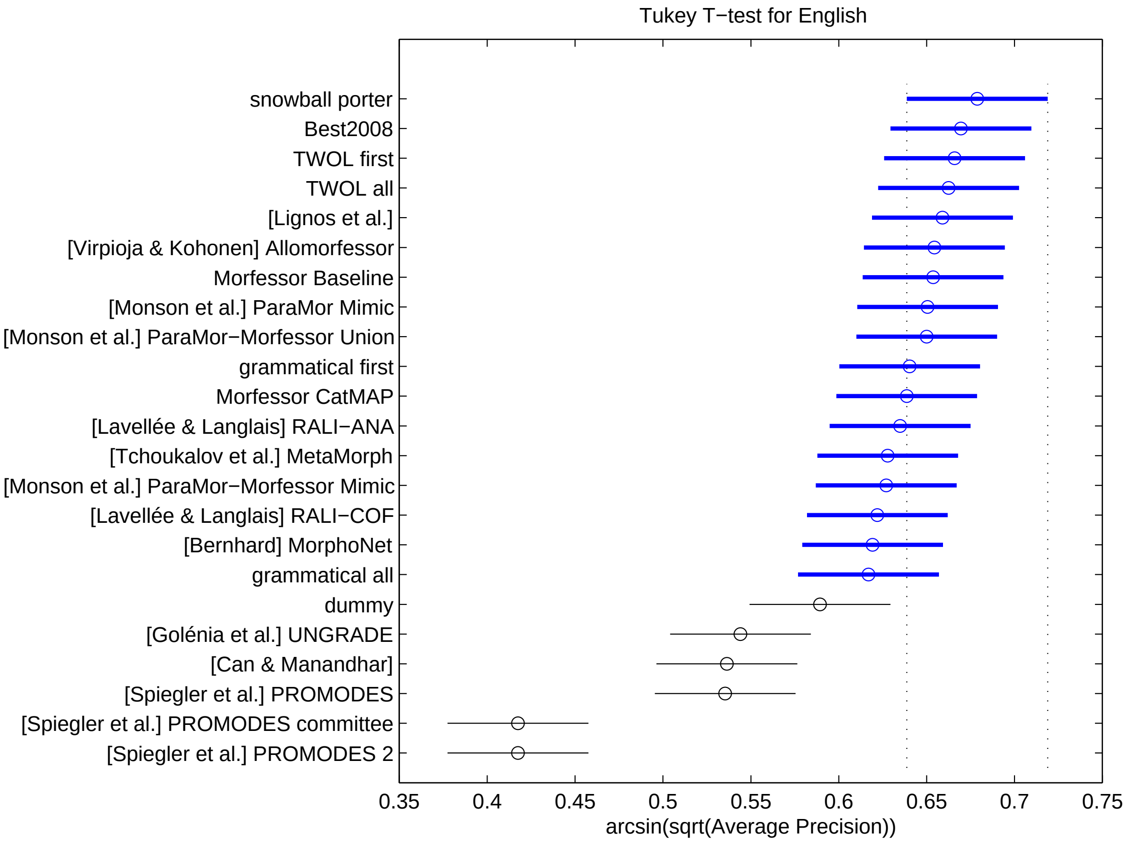

Figure 1: Tukey T-test for English. The "top group" or the submissions that have no significant difference to the best result is highlighted.

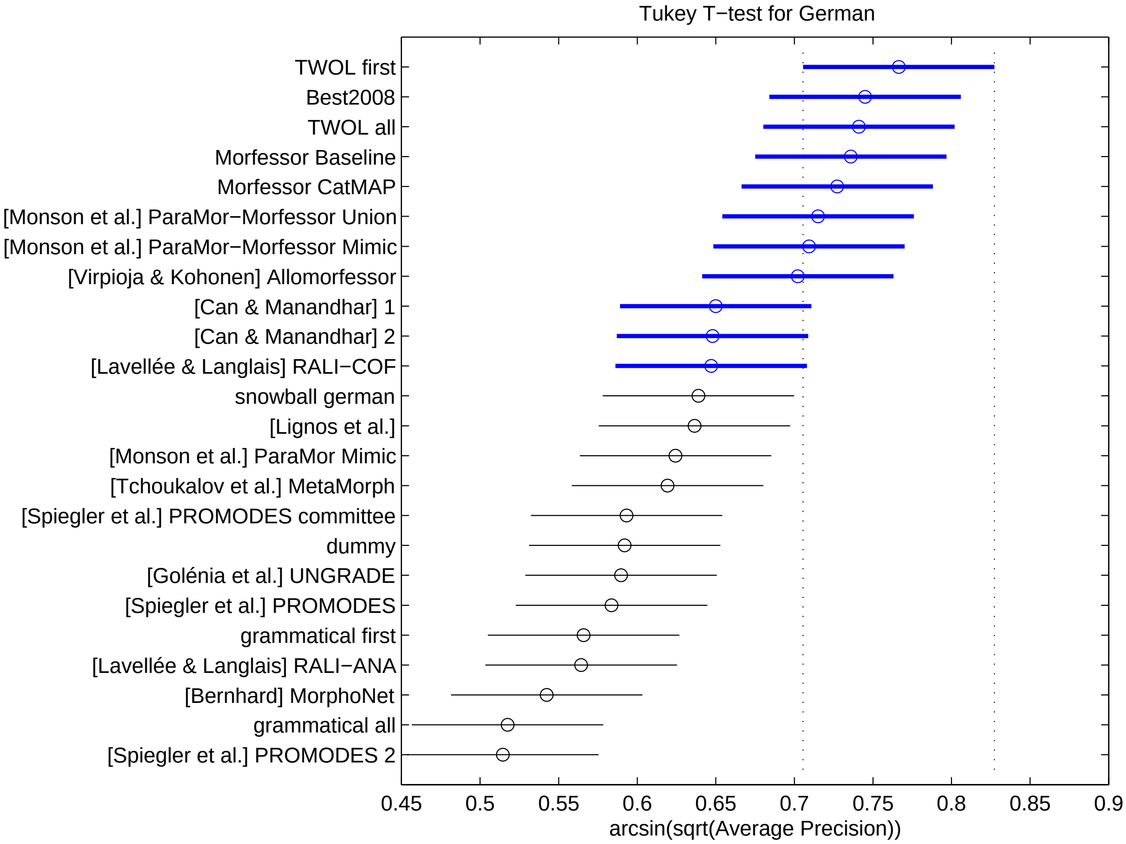

Figure 2: Tukey T-test for German. The "top group" or the submissions that have no significant difference to the best result is highlighted.

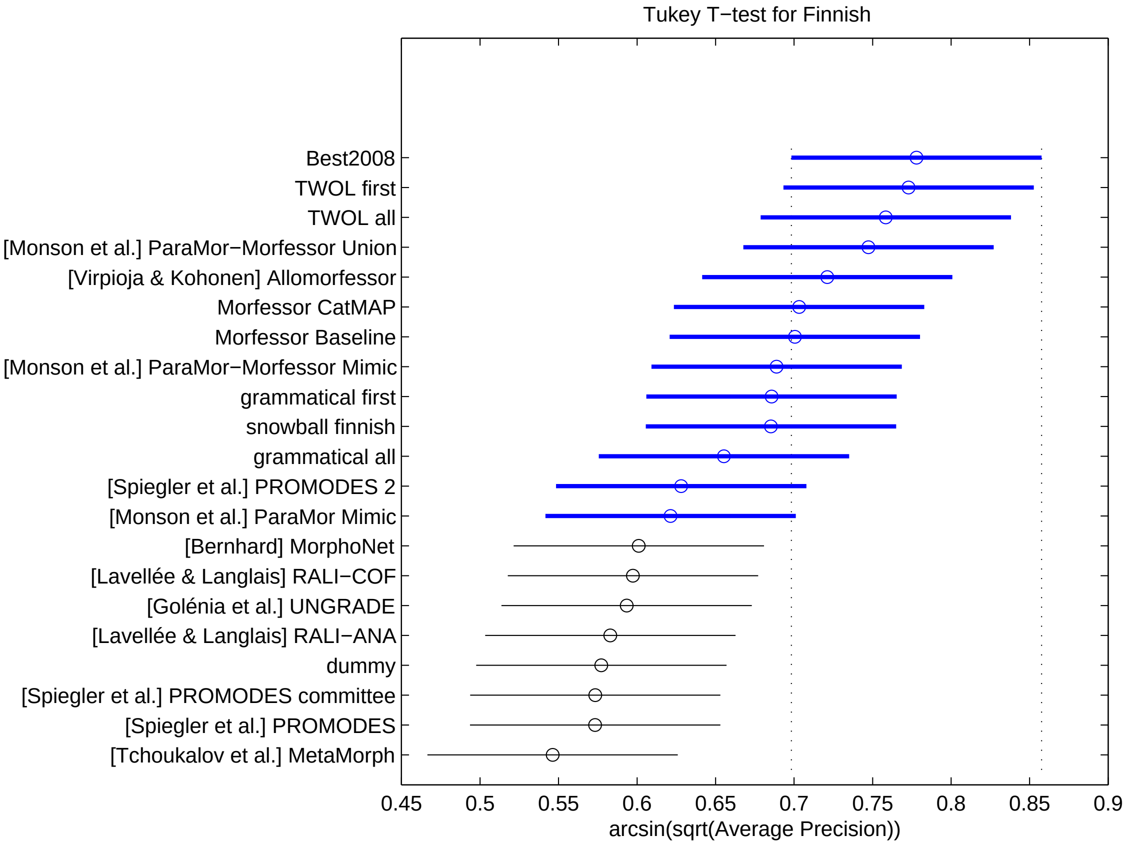

Figure 3: Tukey T-test for Finnish. The “top group” or the submissions that have no significant difference to the best result is highlighted.

and the suggested morphemes give good results when compared to gold standard analysis but do not seem to work well as index terms. On the other hand, "Allomorfessor" and the "Morfessor Baseline"methods performed well in the IR task but were not at the top in the linguistic evaluation where they suffered from low recall. In general, it seems that precision in the Competition 1 evaluation is a better predictor of IR performance than recall or F-measure.

The statistical testing revealed very few significant differences in the IR performance between participants. This is typical for the task. However, we feel that testing the algorithms in a realistic application gives information about the performance of the algorithms that the linguistic comparison can not offer alone.

The participants were offered a chance to access the IR corpus to use the full text context in the unsupervised morpheme analysis. However, this version of task was not attempted by anyone. We are thinking of ways to make this version of task more accessible for competitors as using the context of words seems like a natural way to improve the models. Other future work includes expanding the IR task to new languages like Arabic which pose new kinds of morphological problems.

5 Competition 3 – Statistical Machine Translation

In Competition 3, the morpheme analyses proposed by the participants' algorithm were evaluated in a statistical machine translation (SMT) framework. The translation models were trained to translate from a morphologically complex source language to English. The words of the source language were replaced by their morpheme analyses before training the translation models. The two source languages used in the competition were Finnish and German. Both the input data for the participants' algorithms and training the SMT system were from the proceedings of the European Parliament. The final SMT systems were evaluated by measuring the similarity of the translation results to a human-made reference translation.

5.1 Task and data

As a data set, we used Finnish-English and German-English parts of the European Parliament parallel corpus (release v2) [11]. The participants were given a list of word forms extracted from the corpora, and similarly to the Competitions 1 and 2, they were asked to apply their algorithms to the word list, and return the morphological analyses for the words. It was also possible to use the context information of the words by downloading the full corpus. Furthermore, the data sets from Competitions 1 and 2 were allowed to use for training the morpheme analyses. However, they were used by none of the participants.

For training and testing the SMT systems, the Europarl data sets were divided into three subsets: training set for training the models, development set for tuning the model parameters, and test set for evaluating the translations. For the Finnish-English systems, we had 1 180 603 sentences for training, 2 849 for tuning, and 3 000 for testing. For the German-English systems, we had 1 293 626 sentences for training, 2 665 for tuning, and 3 000 for testing.

5.2 Evaluation

In principle, the evaluation is simple: First, we train a translation system that can translate the morphologically analyzed Finnish or German sentence to English. Then, we use it to translate new sentences, and compare the translation results to the reference translations. If the morphological analysis is good, it reduces the sparsity of the data and helps the translation task. If the analysis contains many errors, they should degrade the translation results. However, a SMT system has many components and parameters that can affect the overall results. Here we describe the full evaluation procedure in detail.

As the SMT models and tools are mainly designed for word-based translations, the results obtained for morpheme-based models are rarely better than the word-based baseline models (see, e.g., [25]). Thus, following the approach in [9], we combined the morpheme-based models to a standard word-based model by generating n-best lists of translation hypotheses from both models, and finding the best overall translation with the Minimum Bayes Risk (MBR) decoding.

5.2.1 Training phrase-based SMT systems

The individual models, including the baseline word-to-word model and the morpheme-to-word models based on the participants' methods, were trained with the open source Moses system [12]. Moses translates sequences of tokens, called phrases, at a time. The decoder finds the most probable hypothesis as a sequence of target language tokens, given a sequence of tokens in source language, a language model, a translation model and possible additional models, such as a reordering model for phrases in the hypothesis.

Training a translation model with Moses includes three main steps: (1) alignment of the tokens in the sentence pairs (2) extracting the phrases from the aligned data, and (3) scoring the extracted phrases. As there are more morphemes than words in a sentence, two limitations affect the results: First, the alignment tool cannot align sentences longer than 100 tokens. Second, the phrases have a maximum length, which we set to be 10 for the morpheme-based models.

The weights of the different components (translation model, language model, etc.) are tuned by maximizing the BLEU score [20] for the development set. Finally, we generated n-best list for the development and test data for the MBR combination. At most 200 distinct hypotheses were generated for each sentence; less if the decoder could not find as many.

5.2.2 Minimum Bayes-Risk decoding for system combination

Minimum Bayes-Risk (MBR) decoding for machine translation [13] selects the translation hypothesis that has the lowest expected risk given the underlying probabilistic model. For loss function L bounded by maximum loss L_{max}, we choose the hypothesis that maximises the conditional expected gain according to the decision rule

$$\hat{E} = \operatorname*{argmax}_{E' \in \mathcal{E}} \sum_{E \in \mathcal{E}} G(E, E') P(E|F), \tag{3}$$

where $G(E, E') = L_{max} - L(E, E')$ is the gain between reference E and hypothesis E' and $P(E|F)$ is the posterior probability of translation. The search is performed over all hypotheses E' in the evidence space $\mathcal{E}$, typically an n-best list or lattice. An appropriate gain function for machine translation is the sentence-level BLEU score [20]. For efficient application to both n-best lists and lattices, our MBR decoder uses an approximation to the sentence-level BLEU score formulated in terms of n-gram posterior probabilities [24]. The contribution of each n-gram w is a constant θ_w multiplied by the number of times w occurs in E' or zero if it does not occur. The decision rule is then

$$\hat{E} = \operatorname*{argmax}_{E' \in \mathcal{E}} \Big\{ \theta_0 |E'| + \sum_{w \in \mathcal{N}} \theta_w \#_w(E') p(w|\mathcal{E}) \Big\}, \tag{4}$$

where $p(w|\mathcal{E})$ is the posterior probability of the n-gram w and $\mathcal{N} = \{w_1, \ldots, w_{|\mathcal{N}|}\}$ denotes the set of all n-grams in the evidence space. The posterior probabilities are computed efficiently using the OpenFst toolkit [2].

We used minimum Bayes-risk system combination [23] to combine n-best list evidence spaces generated by multiple MT systems. The posterior probability of n-gram w in the union of two n-best lists $\mathcal{E}_1$ and $\mathcal{E}_2$ is computed as a linear interpolation of the posterior probabilities according to each individual list:

$$p(w|\mathcal{E}_1 \cup \mathcal{E}_2) = \lambda P(w|\mathcal{E}_1) + (1 - \lambda) P(w|\mathcal{E}_2). \tag{5}$$

The parameter λ determines the weight associated with the output of each translation system and was optimized for BLEU score on the development set.

5.2.3 Evaluation of the translations

For evaluation of the performance of the SMT systems, we applied BLEU scores [20]. BLEU is based on the co-occurrence of n-grams: It counts how many n-grams (for $n = 1, \ldots, 4$) the proposed translation has in common with the reference translations and calculates a score based on this. Although BLEU is a very simplistic method, it usually corresponds well to human evaluations if the compared systems are similar enough. In our case they should be very similar, as the only varying factor is the morphological analysis. In addition to the MBR combinations, we calculated the BLEU scores for all the individual systems.

5.3 Results

Six methods from four groups were included in Competition 3. In addition, Morfessor Baseline and Morfessor Categories-MAP were tested as reference methods. We calculated the BLEU scores both for the individual systems, including a word-based system, and for MBR combination with the word-based system. The results are in Tables 18 and 19.

Between the results from the MBR combinations, only some of the differences are statistically significant. The significances were inspected with paired t-test on ten subsets of the test data. In the Finnish to English task, Morfessor Baseline, Allomorfessor, Morfessor CatMAP and MetaMorph are all significantly better than the rest of the algorithms. Between them, the difference between Allomorfessor and the both Morfessor algorithms is not significant, but Allomorfessor and Morfessor Baseline are significantly better than MetaMorph. The differences between the results of the last four algorithms (MorphoNet and ParaMor:s) are not statistically significant. Neither they are significantly better than the word-based system alone.

In the German to English task, only the results of Morfessor Baseline and Allomorfessor have significant differences to the rest of the systems. Morfessor Baseline is significantly better than any of the others expect Allomorfessor and ParaMor Mimic. Allomorfessor is significantly better than the others expect Morfessor Baseline, ParaMor Mimic, ParaMor-Morfessor Mimic and Morfessor CatMAP. None of the rest of the MBR results is significantly higher than the word-based result.

Table 18: The results of the submitted unsupervised morpheme analyses used in machine translation from **Finnish** (Competition 3).

Author	Method	BLEU
MBR combination with word-based model		
-	Morfessor Baseline	0.2861
Virpioja & Kohonen	Allomorfessor	0.2856
Tchoukalov et al.	MetaMorph	0.2820
-	Morfessor CatMAP	0.2814
Monson et al.	ParaMor-Morfessor Union	0.2784
Bernhard	MorphoNet	0.2779
Monson et al.	ParaMor-Morfessor Mimic	0.2773
Monson et al.	ParaMor Mimic	0.2768
Individual systems		
-	words	0.2764
-	Morfessor Baseline	0.2742
Virpioja & Kohonen	Allomorfessor	0.2717
Tchoukalov et al.	MetaMorph	0.2631
-	Morfessor CatMAP	0.2610
Monson et al.	ParaMor-Morfessor Mimic	0.2347
Monson et al.	ParaMor Mimic	0.2252
Bernhard	MorphoNet	0.2245
Monson et al.	ParaMor-Morfessor Union	0.2223

Table 19: The results of the submitted unsupervised morpheme analyses used in machine translation from **German** (Competition 3).

Author	Method	BLEU
MBR combination with word-based model		
-	Morfessor Baseline	0.3119
Virpioja & Kohonen	Allomorfessor	0.3114
Monson et al.	ParaMor Mimic	0.3086
Monson et al.	ParaMor-Morfessor Union	0.3083
Monson et al.	ParaMor-Morfessor Mimic	0.3081
-	Morfessor CatMAP	0.3080
Tchoukalov et al.	MetaMorph	0.3077
Bernhard	MorphoNet	0.3072
Individual systems		
-	words	0.3063
Virpioja & Kohonen	Allomorfessor	0.3001
-	Morfessor Baseline	0.3000
-	Morfessor CatMAP	0.2901
Tchoukalov et al.	MetaMorph	0.2855
Monson et al.	ParaMor Mimic	0.2854
Monson et al.	ParaMor-Morfessor Mimic	0.2821
Bernhard	MorphoNet	0.2734
Monson et al.	ParaMor-Morfessor Union	0.2729

Overall, the Morfessor family of algorithms performed very well in both translation tasks. Categories-MAP was not as good as Morfessor Baseline or Allomorfessor, which is probably explained by the fact that it segmented words to shorter tokens. Also MetaMorph improved significantly the Finnish translations, but was not as useful in German.

5.4 Discussion

This was the first time that machine translation system was used to evaluate the quality of the morphological analysis. As the SMT tools applied are designed mostly for word-based translations, it was not a surprise that some problems arose.

The word alignment tool used by the Moses system, Giza++, has strict limits on sentence lengths. A sentence cannot be longer than 100 tokens, and neither over 9 times longer or shorter than its sentence pair. Too long sentences are pruned away from the training data. Thus, the algorithms that segmented more, generally got less training data for the translation model. However, the dependency between average tokens per word and the amount of filtered training data was not linear, as seen from the left side of Figure 4. For example, the Morfessor CatMAP system could use much more training data than some of the algorithms that, on average, segmented less. Even without considering the decrease to the amount of training data available, oversegmentation is likely to be detrimental in the task, because it makes, e.g., the word alignment problem more complex.

After MBR combination, the rank of the algorithms was not the same as with the individual systems. The respective scores are plotted in the right side of Figure 4. Especially ParaMor-Morfessor Union system helped the word-based model more than its own BLEU score indicated. However, as the improvements were not statistically significant, the improved rank in the MBR combination may be affected more by just chance.

In addition to making the evaluation more fair to the algorithms that use shorter tokens that the others, future work includes testing TWOL-based morphological analyses or gold standard segmentations in the task.

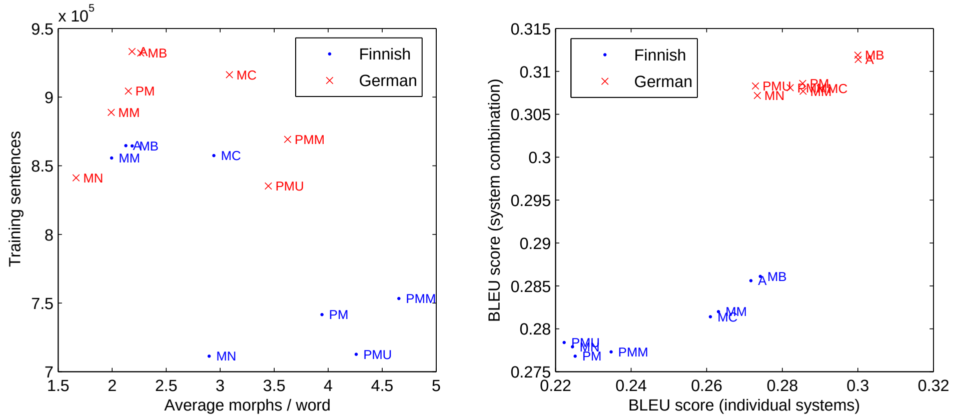

Figure 4: *Left side:* The number of average morphemes per word and the number sentences used for the training the SMT system based on the analysis. *Right side:* BLEU scores of the individual systems and MBR combinations. A = Allomorfessor, MB = Morfessor Baseline, MC = Morfessor CatMAP, MM = MetaMorph, MN = MorphoNet, PM = ParaMor Mimic, PMM = ParaMor-Morfessor Mimic, PMU = ParaMor-Morfessor Union.

6 Conclusion

The Morpho Challenge 2009 was a successful follow-up to our previous Morpho Challenges 2005-2008. Since some of the tasks were unchanged from 2008, the participants of the previous challenges were able to track improvements of their algorithms. It also gave a possibility for the new participants and those who missed the previous deadlines to try more established benchmark tasks. New tasks were introduced for statistical machine translation which offer yet another viewpoint on what is required from morpheme analysis in practical applications.

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 and CLEF who helped us organize this challenge and the challenge workshop. We are most grateful to the University of Leipzig for making the training data resources available to the Challenge, and in particular we thank Stefan Bordag for his kind assistance. We are indebted to Ebru Arisoy for making the Turkish gold standard available to us. We are most grateful to Majdi Sawalha and Eric Atwell from the University of Leeds for making the Arabic data available to the Challenge and for their kind assistance in preparing it to meet the Challenge file formats. We acknowledge also the Computational Linguistics Group at University of Haifa who supplied their tagged Arabic database. This work was supported by the Academy of Finland in the project *Adaptive Informatics*, the graduate schools in Language Technology and Computational Methods of Information Technology, in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, and in part by the IST Programme of the European Community, under the FP7 project EMIME (213845) and PASCAL Network of Excellence. This publication only reflects the authors' views. We acknowledge that access rights to data and other materials are restricted due to other commitments.

References

[1] Eneko Agirre, Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. CLEF 2008: Ad hoc track overview. In *Working Notes for the CLEF 2008 Workshop*, September 2008.

[2] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata (CIAA 2007)*, pages 11–23. Springer Lecture Notes in Computer Science, 2007.

[3] Jeff A. Bilmes and Katrin Kirchhoff. Factored language models and generalized parallel backoff. In *Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 4–6, Edmonton, Canada, 2003.

[4] Mathias Creutz. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30, Philadelphia, Pennsylvania, USA, July 2002.

[5] Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30, 2002.

[6] Mathias Creutz and Krista Lagus. 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, 2005.

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

[8] Mathias Creutz and Krister Linden. Morpheme segmentation gold standards for finnish and english. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004. URL: http://www.cis.hut.fi/projects/morpho/.

[9] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73–76, Boulder, USA, June 2009. Association for Computational Linguistics.

[10] David A. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, New York, NY, USA, 1993. ACM Press.

[11] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of ACL, demonstration session*, Czech Republic, June 2007.

[13] Shankar Kumar and William Byrne. Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 169–176, 2004.

[14] Mikko Kurimo, Mathias Creutz, and Ville Turunen. Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.

[15] Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard – Morpho Challenge 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.

[16] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, and Murat Saraclar. Unsupervised segmentation of words into morphemes - Challenge 2005, an introduction and evaluation report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, Venice, Italy, 2006.

[17] Mikko Kurimo and Ville Turunen. Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[18] Mikko Kurimo and Matti Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[19] Y.-S. Lee. 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, 2004.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[21] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[22] Majdi Sawalha and Eric Atwell. Comparative evaluation of arabic language morphological analysers and stemmers. In *Proceedings of COLING 2008 22nd International Conference on Computational Linguistics*, 2008.

[23] Khe Chai Sim, William J. Byrne, Mark J. F. Gales, Hichem Sahbi, and Phil C. Woodland. Consensus network decoding for statistical machine translation. In *IEEE Conference on Acoustics, Speech and Signal Processing*, 2007.

[24] Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[25] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark, September 2007.

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