

# Overview and Results of Morpho Challenge 2009

Mikko Kurimo<sup>1</sup>, Sami Virpioja<sup>1</sup>, Ville T. Turunen<sup>1</sup>, Graeme W. Blackwood<sup>2</sup>,  
and William Byrne<sup>2</sup>

<sup>1</sup> Adaptive Informatics Research Centre, Helsinki University of Technology,  
P.O.Box 5400, FIN-02015 TKK, Finland

<http://www.cis.hut.fi/morphochallenge2009/>

<sup>2</sup> Cambridge University Engineering Department  
Trumpington Street, Cambridge CB2 1PZ, U.K.

**Abstract.** The goal of Morpho Challenge 2009 was to evaluate unsupervised algorithms that provide morpheme analyses for words in different languages and 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 were: Finnish, Turkish, German, English and Arabic. This paper describes the tasks, evaluation methods, and obtained results. The Morpho Challenge was part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF.

## 1 Introduction

Unsupervised morpheme analysis is one of the important but unsolved tasks in computational linguistics and its applications, such as speech recognition (ASR) [1, 2], information retrieval (IR) [3, 4] and statistical machine translation (SMT) [5, 6]. 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 of 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*), IR (*Competition 2*) and SMT (*Competition 3*) tasks. The IR experiments contain CLEF tasks, where the all the words in the queries and text corpus were replaced by their morpheme analyses. In SMT experiments identical translation 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 for Morpho Challenge and are based on recent work of the organizers in morpheme based machine translation [6, 7].

## 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. The short acronyms of max 8 characters are used in the result tables throughout the paper.

Authors, Affiliations:	Algorithm name [Acronym]	
D. Bernhard, TU Darmstadt, D:	MorphoNet	[MorphNet]
B. Can & S. Manandhar, U. York, UK:	1 [CanMan1], 2	[CanMan2]
B. Golénia et al., U. Bristol, UK:	UNGRADE	[Ungrade]
J-F. Lavallée & P. Langlais, U. Montreal, CA:	RALI-ANA	[Rali-ana],
	RALI-COF	[Rali-cof]
C. Lignos et al., U. Penn. & Arizona, USA:	-	[Lignos]
C. Monson et al., Oregon Health & Sc. U., USA:	ParaMor Mimic	[P-Mimic],
	ParaMor-Morfessor Mimic [PM-Mimic], ParaMor-Morfessor Union	[PM-Union]
S. Spiegler et al., U. Bristol, UK:	PROMODES	[Prom-1],
	PROMODES 2 [Prom-2], PROMODES committee	[Prom-com]
T. Tchoukalov et al., U. Stanford & OHSU, USA:	MetaMorph	[MetaMorf]
S. Virpioja & O. Kohonen, Helsinki U. Tech., FI:	Allomorfessor	[Allomorf]

Statistics of the output of the submitted algorithms are briefly presented in Table 3 for English. The corresponding data for each of the languages is presented in [8]. The average amount of analyses per word is shown in the column

**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-NV	ARA-V	ENG	FIN	GER	TUR	C2	C3
MorphNet	X	X	X	X	X	X	X	X
CanMan1	-	-	X	-	X	X	-	-
CanMan2	-	-	-	-	X	X	-	-
Ungrade	X	X	X	X	X	X	-	-
Rali-ana	X	X	X	X	X	X	-	-
Rali-cof	X	X	X	X	X	X	-	-
Lignos	-	-	X	-	X	-	-	-
P-Mimic	X	X	X	X	X	X	X	X
PM-Mimic	X	X	X	X	X	X	X	X
PM-Union	X	X	X	X	X	X	X	X
Prom-1	X	X	X	X	X	X	-	-
Prom-2	X	X	X	X	X	X	-	-
Prom-com	X	X	X	X	X	X	-	-
MetaMorf	X	X	X	X	X	X	-	X
Allomorf	X	X	X	X	X	X	X	X
Total	12	12	14	12	15	14	5	6

“#analyses”. It is interesting that in contrary to previous years, now all algorithms ended up mostly suggesting only one analysis per word. From the column “#morphs” we see the average amount of morphemes per analysis, which reflects the level of details the algorithm provides. The total amount of morpheme types in the lexicon is given in the column “#types”.

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, CatMAP ” for short) developed at Helsinki University of Technology [9]. Analysis by the original Morfessor [10, 11] (or here “Morfessor Baseline, MorfBase”), 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

**Table 3.** Statistics and example morpheme analyses in **English**. #analyses is the average amount of analyses per word (separated by a comma), #morphs the average amount of morphemes per analysis (separated by a space), and #types the total amount of morpheme types in the lexicon.

Algorithm	#analyses	#morphs	#types	example analysis
MorphNet	1	1.75	211439	vulnerabilty_ies
CanMan	1	2.09	150097	vulner_abilities
Ungrade	1	3.87	123634	vulnerabilities
Rali-ana	1	2.10	166826	vulner_abiliti_es
Rali-cof	1	1.91	145733	vulnerability_ies
Lignos	1	1.74	198546	VULNERABILITY_+(ies)
P-Mimic	1	3.03	188716	vulner_+a_+bilit_+ie_+s
PM-Mimic	1	2.96	166310	vulner_+a_+bilit_+ies
PM-Union	1	2.87	120148	vulner_a_+bilit_+ies
Prom-1	1	3.28	107111	vul_nerabilitie_s
Prom-2	1	3.63	47456	v_ul_nera_b_ili_ties
Prom-com	1	3.63	47456	v_ul_nera_b_ili_ties
MetaMorf	1	1.58	241013	vulnerabiliti_es
Allomorf	1	2.59	23741	vulnerability_ies
MorfBase	1	2.31	40293	vulner_abilities
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

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<sup>3</sup> 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) [12]. 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 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, only morpheme analysis submitted in Roman script, was evaluated.

<sup>3</sup> <http://corpora.informatik.uni-leipzig.de/>



fixed from the scripts and points were now measured as one per word, not one per word pair.

Because the morpheme analysis candidates are achieved by unsupervised learning, the morpheme 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. The proportion of morpheme sharing word pairs in the participant's sample that really has a morpheme in common according to the gold standard is called the *precision*. Correspondingly, the proportion of morpheme sharing word pairs in the gold standard sample that also exist in the participant's submission is called the *recall*.

In practise, the 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. 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.

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 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.

The 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:

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

### 3.3 Results

**Table 5.** The morpheme analyses compared to the gold standard in **non-vowelized and vowelized Arabic** (Competition 1). The numbers are in %.

Method	Non-vowelized Arabic			Method	Vowelized Arabic		
	Precision	Recall	F-measure		Precision	Recall	F-measure
letters	70.48	53.51	60.83	letters	50.56	84.08	63.15
Prom-2	76.96	37.02	50.00	Prom-2	63.00	59.07	60.97
Prom-com	77.06	36.96	49.96	Prom-com	68.32	47.97	56.36
Prom-1	81.10	20.57	32.82	Ungrade	72.15	43.61	54.36
Ungrade	83.48	15.95	26.78	Prom-1	74.85	35.00	47.70
Allomorf	91.62	6.59	12.30	MorfBase	86.87	4.90	9.28
MorfBase	91.77	6.44	12.03	PM-Union	91.61	4.41	8.42
MorphNet	90.49	4.95	9.39	Allomorf	88.28	4.37	8.33
PM-Union	93.72	4.81	9.14	PM-Mimic	93.62	3.23	6.24
PM-Mimic	93.76	4.55	8.67	MorphNet	92.52	2.91	5.65
Rali-ana	92.40	4.40	8.41	MetaMorf	88.78	2.89	5.59
MetaMorf	95.05	2.72	5.29	Rali-ana	91.30	2.83	5.49
P-Mimic	91.29	2.56	4.97	P-Mimic	91.36	1.85	3.63
Rali-cof	94.56	2.13	4.18	Rali-cof	95.09	1.50	2.95

The results of the Competition 1 are presented in Tables 5 and 6. In three languages, Turkish, Finnish and German, the algorithms with clearly highest F-measures 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 [14] [PM-2008], [ParaMor] and 2007 [13] [Bernhard2], [Bordag5a]. From Morpho Challenge 2008, the best method “Paramor + Morfessor 2008” [PM-2008] would have also scored highest in 2009. However, this 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

**Table 6.** The morpheme analyses compared to the gold standard in %. The results below the line are by the winners of the previous Morpho Challenges.

Method	English			Method	German		
	Precision	Recall	F-measure		Precision	Recall	F-measure
Allomorf	68.98	56.82	62.31	PM-Union	52.53	60.27	56.14
MorfBase	74.93	49.81	59.84	PM-Mimic	51.07	57.79	54.22
PM-Union	55.68	62.33	58.82	CatMAP	71.08	38.92	50.30
Lignos	83.49	45.00	58.48	P-Mimic	50.81	47.68	49.20
P-Mimic	53.13	59.01	55.91	CanMan2	57.67	42.67	49.05
MorphNet	65.08	47.82	55.13	Rali-cof	67.53	34.38	45.57
PM-Mimic	54.80	60.17	57.36	Prom-2	36.11	50.52	42.12
Rali-cof	68.32	46.45	55.30	Allomorf	77.78	28.83	42.07
CanMan1	58.52	44.82	50.76	MorphNet	67.41	30.19	41.71
CatMAP	84.75	35.97	50.50	Prom-1	49.88	33.95	40.40
Prom-1	36.20	64.81	46.46	Prom-com	48.48	34.61	40.39
Rali-ana	64.61	33.48	44.10	MorfBase	81.70	22.98	35.87
Prom-2	32.24	61.10	42.21	Lignos	78.90	21.35	33.61
Prom-com	32.24	61.10	42.21	Ungrade	39.02	29.25	33.44
MetaMorf	68.41	27.55	39.29	MetaMorf	39.59	19.81	26.40
Ungrade	28.29	51.74	36.58	CanMan1	73.16	15.27	25.27
letters	3.82	99.88	7.35	Rali-ana	61.39	15.34	24.55
				letters	2.79	99.92	5.43
PM-2008	69.59	65.57	67.52	PM-2008	64.06	61.52	62.76
ParaMor	63.32	51.96	57.08	ParaMor	70.73	38.82	50.13
Bernhard2	67.42	65.11	66.24	Bernhard2	54.02	60.77	57.20
Method	Finnish			Method	Turkish		
	Precision	Recall	F-measure		Precision	Recall	F-measure
PM-Union	47.89	50.98	49.39	PM-Mimic	48.07	60.39	53.53
PM-Mimic	51.75	45.42	48.38	PM-Union	47.25	60.01	52.88
CatMAP	79.01	31.08	44.61	P-Mimic	49.54	54.77	52.02
Prom-com	41.20	48.22	44.44	Rali-cof	48.43	44.54	46.40
P-Mimic	47.15	40.50	43.57	CatMAP	79.38	31.88	45.49
Prom-2	33.51	61.32	43.34	Prom-2	35.36	58.70	44.14
Prom-1	35.86	51.41	42.25	Prom-1	32.22	66.42	43.39
Rali-cof	74.76	26.20	38.81	MorphNet	61.75	30.90	41.19
Ungrade	40.78	33.02	36.49	CanMan2	41.39	38.13	39.70
MorphNet	63.35	22.62	33.34	Prom-com	55.30	28.35	37.48
Allomorf	86.51	19.96	32.44	Ungrade	46.67	30.16	36.64
MorfBase	89.41	15.73	26.75	MetaMorf	39.14	29.45	33.61
MetaMorf	37.17	15.15	21.53	Allomorf	85.89	19.53	31.82
Rali-ana	60.06	10.33	17.63	MorfBase	89.68	17.78	29.67
letters	5.17	99.89	9.83	Rali-ana	69.52	12.85	21.69
				letters	8.66	99.13	15.93
				CanMan1	73.03	8.89	15.86
PM-2008	65.21	50.43	56.87	PM-2008	66.78	57.97	62.07
ParaMor	49.97	37.64	42.93	ParaMor	57.35	45.75	50.90
Bernhard2	63.92	44.48	52.45	Bordag5a	81.06	23.51	36.45

scores. However, exploiting this property of the evaluation measure is not a very interesting approach.

Excluding “Paramor + Morfessor 2008”, 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 [Bernhard2] 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 IR tasks with three languages: English, German and Finnish. The tasks and corpora were the same as in 2007 [4] and 2008 [15]. In the evaluation, words occurring in the corpus and in the queries were replaced by the morpheme segmentations submitted by the participants. 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.

### 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 information need. To evaluate 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)<sup>4</sup> 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)

---

<sup>4</sup> <http://www.clef-campaign.org/>

- 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 participants' submissions were compared against a number of relevant reference methods which were the same as used in Morpho Challenge 2008 [15]. Like the participants' methods, Morfessor baseline [MorfBase] [16, 11] and Morfessor Categories-MAP [CatMAP] [9] are unsupervised algorithms. Also evaluated were a commercial word normalization tool [TWOL] and the rule-based grammatical morpheme analyses [gram] based on the linguistic gold standards [17]. These methods have the benefit of language specific linguistic knowledge embedded in them. Because some words may have several alternative interpretations two versions of these references cases were given: either all alternatives were used (e.g. [TWOL-all]) or only the first one (e.g. [TWOL-1]). Traditional language-dependent stemming approaches based on the Snowball libstemmer library [StemEng], [StemGer] and [StemFin] as well as using the words without any processing were also tested [dummy].

In each task the best algorithm in 2008, i.e. the one that provided the highest average precision ([PM-2008] and [McNamee4]) can be used as a reference, too, because 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<sup>5</sup> version 4.4. The popular Okapi BM25 ranking function was used. In the 2007 challenge [4], 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

<sup>5</sup> <http://www.lemurproject.org/>

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.

Table 7 shows the obtained MAP values for the submissions in English, German and Finnish. 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 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 [18] and CLEF [19].

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:

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

**Table 7.** The obtained mean average precision (MAP) in the IR tasks. Asterisk (\*) denotes submissions that did not include segmentations for Competition 2 and were evaluated by using the shorter Competition 1 word list. The results below the line are statistically significantly different from the best result of that language.

Method	English	Method	German	Method	Finnish
StemEng	0.4081	TWOL-1	0.4885	TWOL-1	0.4976
PM-2008	0.3989	TWOL-all	0.4743	McNamee4	0.4918
TWOL-1	0.3957	PM-2008	0.4734	TWOL-all	0.4845
TWOL-all	0.3922	MorfBase	0.4656	PM-Union	0.4713
Lignos	0.3890*	CatMAP	0.4642	Allomorf	0.4601
MorfBase	0.3861	PM-Mimic	0.4490	CatMAP	0.4441
Allomorf	0.3852	PM-Union	0.4478	MorfBase	0.4425
P-Mimic	0.3822	Allomorf	0.4388	gram-1	0.4312
PM-Union	0.3811	CanMan1	0.4006*	PM-Mimic	0.4294
gram-1	0.3734	Rali-cof	0.3965*	StemFin	0.4275
CatMAP	0.3713	CanMan2	0.3952*	gram-all	0.4090
Rali-ana	0.3707*			Prom-2	0.3857*
PM-Mimic	0.3649			P-Mimic	0.3819
MetaMorf	0.3623*				
Rali-cof	0.3616*				
MorphNet	0.3560				
gram-all	0.3542				
dummy	0.3293	StemGer	0.3865	Rali-cof	0.3740*
Ungrade	0.2996*	Lignos	0.3863*	MorphNet	0.3668
CanMan1	0.2940*	P-Mimic	0.3757	Ungrade	0.3636*
Prom-1	0.2917*	MetaMorf	0.3752*	Rali-ana	0.3595*
Prom-2	0.2066*	Prom-com	0.3634*	dummy	0.3519
Prom-com	0.2066*	dummy	0.3509	Prom-com	0.3492*
		Ungrade	0.3496*	Prom-1	0.3392*
		Prom-1	0.3484*	MetaMorf	0.3289*
		gram-1	0.3353		
		Rali-ana	0.3284*		
		MorphNet	0.3167		
		gram-all	0.3014		
		Prom-2	0.2997*		

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.

Based on the confidence test results a horizontal line is drawn in Table 7 at the point where all the methods below it are significantly different from the best result, and the “top group” above it are those that have no significant difference to the best result of each language. Further analysis of the confidence test results are in [8]. 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 [19] 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 IR. 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 [4], 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 exceptions 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 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. Although using the context of words seems a natural way to improve the models none of the participants have attempted this. 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 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) [20]. 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., [6]). Thus, following the approach in [7], 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.

**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 [21]. 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 [22] 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.

**Minimum Bayes-Risk decoding for system combination** Minimum Bayes-Risk (MBR) decoding for machine translation [23] 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

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

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 [22]. 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  $\theta_w$  multiplied by the number of times  $w$  occurs in  $E'$  or zero if it does not occur. The decision rule is then

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

where  $p(w|\mathcal{E})$  is the posterior probability of the  $n$ -gram  $w$  and  $\mathcal{N} = \{w_1, \dots, w_{|\mathcal{N}|}\}$  denotes the set of all  $n$ -grams in the evidence space. The posterior probabilities are computed efficiently using the OpenFst toolkit [25].

We used minimum Bayes-risk system combination [26] 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:

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

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

**Evaluation of the translations** For evaluation of the performance of the SMT systems, we applied BLEU scores [22]. BLEU is based on the co-occurrence of  $n$ -grams: It counts how many  $n$ -grams (for  $n = 1, \dots, 4$ ) 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 [MorfBase], Morfessor Categories-MAP [CatMAP] and grammatical morphemes [gram-1] were tested as reference methods. We calculated the BLEU scores both for the individual systems, including a word-based system [words], and for MBR combination with the word-based system. The results are in Table 8.

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 except Allomorfessor and ParaMor Mimic. Allomorfessor is significantly better than the others except 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 8.** The BLEU results of the submitted unsupervised morpheme analyses used in SMT from **Finnish and German** for both Individual systems and MBR combination with word-based models (Competition 3).

Finnish-English				German-English			
Method	Comb.	Method	Indiv.	Method	Comb.	Method	Indiv.
MorfBase	0.2861	MorfBase	0.2742	MorfBase	0.3119	Allomorf	0.3001
Allomorf	0.2856	Allomorf	0.2717	Allomorf	0.3114	MorfBase	0.3000
gram-1	0.2821	MetaMorf	0.2631	gram-1	0.3103	CatMAP	0.2901
MetaMorf	0.2820	CatMAP	0.2610	P-Mimic	0.3086	gram-1	0.2873
CatMAP	0.2814	gram-1	0.2580	PM-Union	0.3083	MetaMorf	0.2855
PM-Union	0.2784	PM-Mimic	0.2347	PM-Mimic	0.3081	P-Mimic	0.2854
MorphNet	0.2779	P-Mimic	0.2252	CatMAP	0.3080	PM-Mimic	0.2821
PM-Mimic	0.2773	MorphNet	0.2245	MetaMorf	0.3077	MorphNet	0.2734
P-Mimic	0.2768	PM-Union	0.2223	MorphNet	0.3072	PM-Union	0.2729
		words	0.2764			words	0.3063

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 a SMT 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. 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. However, this sentence length restriction should be solved for the future evaluations.

After MBR combination, the rank of the algorithms was not the same as with the individual systems. 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.

## 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 SMT which offer yet another viewpoint on what is required from morpheme analysis in practical applications.

The various evaluation results indicate the benefit of utilizing real-world applications for studying the morpheme analysis methods. Some algorithms that succeeded relatively well in imitating the grammatical morphemes did not perform as well in applications as others that differed more from the grammatical ones. Although the mutual performance differences of various algorithms in applications are often small, it seems that different applications may favor different kinds of morpheme, and thus, proposing the overall best morphemes is difficult.

## Acknowledgments

We thank all the participants for their submissions and enthusiasm and the organizers of the PASCAL Challenge Program and CLEF who helped us to organize this challenge and its workshop. We are grateful to the University of Leipzig, University of Leeds, Computational Linguistics Group at University of Haifa, Stefan Bordag, Ebru Arisoy, Majdi Sawalha, Eric Atwell, and Mathias Creutz for making the data and gold standards in various languages available to the Challenge. 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. Bilmes, J.A., Kirchoff, K.: 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), Edmonton, Canada (2003) 4–6
2. Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E., Saraclar, M.: 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)
3. Ziemann, Y., Bleich, H.: 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)
4. Kurimo, M., Creutz, M., Turunen, V.: Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2007. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
5. Lee, Y.S.: 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)
6. Virpioja, S., Väyrynen, J.J., Creutz, M., Sadeniemi, M.: Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In: Proceedings of the Machine Translation Summit XI, Copenhagen, Denmark (September 2007) 491–498
7. de Gispert, A., Virpioja, S., Kurimo, M., Byrne, W.: 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, Boulder, USA, Association for Computational Linguistics (June 2009) 73–76
8. Kurimo, M., Virpioja, S., Turunen, V.T., Blackwood, G.W., Byrne, W.: Overview and results of Morpho Challenge 2009. In: Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)
9. Creutz, M., Lagus, K.: Inducing the morphological lexicon of a natural language from unannotated text. In: Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Finland (2005) 106–113
10. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02. (2002) 21–30
11. Creutz, M., Lagus, K.: 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/>.
12. Sawalha, M., Atwell, E.: Comparative evaluation of arabic language morphological analysers and stemmers. In: Proceedings of COLING 2008 22nd International Conference on Computational Linguistics. (2008)

13. Kurimo, M., Creutz, M., Varjokallio, M.: 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)
14. Kurimo, M., Varjokallio, M.: 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)
15. Kurimo, M., Turunen, V.: Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark (2008)
16. Creutz, M.: Unsupervised discovery of morphemes. In: Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, Philadelphia, Pennsylvania, USA (July 2002) 21–30
17. Creutz, M., Linden, K.: 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/>.
18. Hull, D.A.: 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, New York, NY, USA, ACM Press (1993) 329–338
19. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2008: Ad hoc track overview. In: Working Notes for the CLEF 2008 Workshop. (September 2008)
20. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the 10th Machine Translation Summit, Phuket, Thailand (2005) 79–86
21. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Annual Meeting of ACL, demonstration session, Czech Republic (June 2007)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02), Morristown, NJ, USA, Association for Computational Linguistics (2002) 311–318
23. Kumar, S., Byrne, W.: 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. (2004) 169–176
24. Tromble, R., Kumar, S., Och, F., Macherey, W.: Lattice Minimum Bayes-Risk decoding for statistical machine translation. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, Association for Computational Linguistics (October 2008) 620–629
25. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., Mohri, M.: 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), Springer Lecture Notes in Computer Science (2007) 11–23
26. Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., Woodland, P.C.: Consensus network decoding for statistical machine translation. In: IEEE Conference on Acoustics, Speech and Signal Processing. (2007)