## Datasets

There are a number of data files involved in this challenge. Each type of file is available for each language.

***NEW (2009-02-25):*** *All random word pairs files have been updated so that they correspond to the new evaluation scripts. In addition, small modifications have been made also to the Arabic word lists and gold standard samples.*

### Word list (input)

First and foremost, there is a list of word forms. The words have been extracted from a text corpus, and each word in the list is preceded by its frequency in the corpus used.

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

```
...
1 barefoot's
2 barefooted
6699 feet
653 flies
2939 flying
1782 foot
64 footprints
...
```

### Result file (output, i.e., what to submit)

The participants' task is to return a list containing exactly the same words as in the input, with morpheme analyses provided for each word. The list returned shall not contain the word frequency information.

A submission for the above English words may look like this:

```
...
barefoot's      BARE FOOT +GEN
barefooted      BARE FOOT +PAST
feet            FOOT +PL
flies           FLY_N +PL, FLY_V +3SG
flying          FLY_V +PCP1
foot            FOOT
footprints      FOOT PRINT +PL
...
```

There are a number of things to note about the result file: Each line of the file contains a word (e.g., "feet") separated from its analysis (e.g., "FOOT +PL") by one TAB character. The word needs to look exactly as it does in the input; no capitalization or change of character encoding is allowed. The analysis contains morpheme labels separated using space. The order in which the labels appear does not matter; e.g., "FOOT +PL" is equivalent to "+PL FOOT". The labels are arbitrary: e.g., instead of using "FOOT" you might use "morpheme784" and instead of "+PL" you might use "morpheme2". However, we strongly recommend you to use intuitive labels, when possible, since they make it easier for anyone to get an idea of the quality of the result by looking at it.

If a word has several interpretations, all interpretations should be supplied: e.g., the word "flies" may be the plural form of the noun "fly" (insect) or the third person singular present tense form of the verb "to fly". The alternative analyses must be separated using a comma, as in: "FLY_N +PL, FLY_V +3SG". The existence of alternative analyses makes the task challenging, and we leave it to the participants to decide how much effort they will put into this aspect of the task. In English, for instance, in order to get a perfect score, it would be necessary to distinguish the different functions of the ending "-s" (plural or person ending) as well as the different parts-of-speech of the stem "fly" (noun or verb). As the results will be evaluated against reference analyses (our so-called gold standard), it is worth **reading about the guiding principles used when constructing the gold standard**.

As far as we understand, you can use any characters in your morpheme labels except whitespace and comma (,). However, we cannot guarantee that the evaluation scripts will work properly, if your labels contain some "strange" characters.

### Text corpus for English, Finnish, German and Turkish

The word list (input data) has been constructed by collecting word forms occurring in a text corpus. The text corpora have been obtained from the ***Wortschatz* collection** at the University of Leipzig (Germany). We used the plain text files (`sentences.txt` for each language); the corpus sizes are 3 million sentences for English, Finnish and German, and 1 million sentences for Turkish. For English, Finnish and Turkish we use preliminary corpora, which have not yet been released publicly at the Wortschatz site. The corpora have been preprocessed for the Morpho Challenge (tokenized, lower-cased, some conversion of character encodings).

If the participants like to do so, they can use the corpora in order to get information about the context in which the different words occur.

We are most grateful to the University of Leipzig for making these resources available to the Challenge, and in particular we thank Stefan Bordag for his kind assistance.

### Text corpus for Arabic

**NEW:** This year we try a different data set, the Quran, which is somewhat smaller (only 78K words), but has also a vowelized version (as well as the unvowelized one). The text data has also been made available.
In Arabic, the participants can try to analyze the vowelized words or the unvowelized, or both. They will be evaluated separately against the vowelized or the unvowelized gold standard analysis, respectively.
For all Arabic data, the Arabic writing script are provided as well as the Roman script (Buckwalter transliteration). However, we can only evaluate morpheme analysis submitted in Roman script, sorry.

We are most grateful to Majdi Sawalha and Eric Atwell from the University of Leeds for making this data available to the Challenge and for their kind assistance in preparing it to meet the Challenge file formats.
Sawalha, Majdi; Atwell, Eric. 2008. Comparative evaluation of Arabic language morphological analysers and stemmers. in: Proceedings of COLING 2008 22nd International Conference on Computational Linguistics. **[PDF]**
We acknowledge also the Computational Linguistics Group at University of Haifa who supplied their **tagged database**.

## Gold standard morpheme analyses

The desired "correct" analyses for a random sample of circa 500 words are supplied for each language. These samples can be used for visual inspection and as a *development test set* (in order to get a rough estimate of the performance of the participants' morpheme-analyzing algorithm).

The format of the gold standard file is exactly the same as that of the **result file** to be submitted. That is, each line contains a word and its analysis. The word is separated from the analysis by a TAB character. Morpheme labels in the analysis are separated from each other by a space character. For some words there are multiple correct analyses. These alternative analyses are separated by a comma (,). Examples:

| Language | Examples | |
|---|---|---|
| English | `baby-sitters` | `baby_N sit_V er_s +PL` |
| | `indoctrinated` | `in_p doctrine_N ate_s +PAST` |
| Finnish | `linuxiin` | `linux_N +ILL` |
| | `makaronia` | `makaroni_N +PTV` |
| German | `choreographische` | `choreographie_N isch +ADJ-e` |
| | `zurueckzubehalten` | `zurueck_B zu be halt_V +INF` |
| Turkish | `kontrole` | `kontrol +DAT` |
| | `popUlerliGini` | `popUler +DER_lHg +POS2S +ACC, popUler +DER_lHg +POS3 +ACC3` |
| Arabic vowelized | `Al>aroDi` | `'rD faEl 'arD +Noun +Triptotic +Sg +Fem +Gen +Def` |
| Arabic non-vowelized | `Al>rD` | `'rD fEl 'rD +Noun +Triptotic +Sg +Fem +Gen +Def` |

The English and German gold standards are based on the **CELEX data base**. The Finnish gold standard is based on the two-level morphology analyzer FINTWOL from **Lingsoft, Inc.** The Turkish gold-standard analyses have been obtained from a morphological parser developed at **Boğaziçi University**; it is based on Oflazer's finite-state machines, with a number of changes. We are indebted to Ebru Arısoy for making the Turkish gold standard available to us.

For Arabic the gold standard has in each line; the word, the root, the pattern and then the morphological and part-of-speech analysis.

The morphological analyses are *morpheme* analyses. This means that only grammatical categories that are realized as morphemes are included. For instance, for none of the languages will you find a singular morpheme for nouns or a present-tense morpheme for verbs, because these grammatical categories do not alter or add anything to the word form, in contrast to, e.g., the plural form of a noun (house vs. house+*s*), or the past tense of verbs (help vs. help+*ed*, come vs. c*a*me).

The morpheme labels that correspond to inflectional (and sometimes also derivational) *affixes* have been marked with an initial plus sign (e.g., +PL, +PAST). This is due to a feature of the evaluation script: in addition to the overall performance statistics, evaluation measures are also computed separately for the labels starting with a plus sign and those without an initial plus sign. It is thus possible to make an approximate assessment of how accurately affixes are analyzed vs. non-affixes (mostly stems). If you use the same naming convention when labeling the morphemes proposed by your algorithm, this kind of statistics will be available for your output (see the **evaluation page** for more information).

The morpheme labels that have not been marked as affixes (no initial plus sign) are typically stems. These labels consist of an intuitive string, usually followed by an underscore character (_) and a part-of-speech tag, e.g., "baby_N", "sit_V". In many cases, especially in English, the same morpheme can function as different parts-of-speech; e.g., the English word "force" can be a noun or a verb. In the majority of these cases, however, if there is only a difference in syntax (and not in meaning), the morpheme has been labeled as *either* a noun or a verb, throughout. For instance, the "original" part-of-speech of "force" is a noun, and consequently both noun and verb inflections of "force" contain the morpheme "force_N":

```
force     force_N
force's   force_N GEN
forced    force_N +PAST
forces    force_N +3SG, force_N +PL
forcing   force_N +PCP1
```

Thus, there is not really a need for your algorithm to distinguish between different meanings or syntactic roles of the discovered stem morphemes. However, in some rare cases, if the meanings of the different parts-of-speech do differ clearly, there are two variants, e.g., "train_N" (vehicle), "train_V" (to teach), "fly_N" (insect), "fly_V" (to move through the air). But again, if there are ambiguous meanings *within* the same part-of-speech, these are *not* marked in any way, e.g., "fan_N" (device for producing a current of air) vs. "fan_N" (admirer). This notation is a consequence of using CELEX and FINTWOL as the sources for our gold standards. We could have removed the part-of-speech tags, but we decided to leave them there, since they carry useful information without significantly making the task more difficult. There are no part-of-speech tags in the Turkish gold standard.

## Random word pairs file

If you want to carry out a small-scale evaluation yourself using the gold standard sample, you need to download a randomly generated so-called *word pairs file* for each language to be tested. Read more about this on the **evaluation page**.

## Character encoding

In the source data used for the different languages, there is variation in how accurately certain distinctions are made when letters are rendered. This makes it hard to apply a unified character encoding scheme for all the languages (such as UTF-8). Thus, the following encodings have been used, in which all letters are encoded as one-byte (8-bit) characters:

English
: Standard text. All words are lower-cased, also proper names.

Finnish

ISO Latin 1 (ISO 8859-1). The Scandinavian special letters å, ä, ö (as well as other letters occuring in loan words, e.g., ü, é, à) are rendered as one-byte characters. All words are lower-cased, also proper names.

German

Standard text. All words are lower-cased, also all nouns. The German umlaut letters are rendered as the corresponding non-umlaut letter followed by "e", e.g., "laender" (Länder), "koennte" (könnte), "fuer" (für). Double-s is rendered as "ss", e.g., "strasse" (Straße). This coarse encoding is due to the fact that CELEX, the source for the morphological gold standard, utilizes this scheme. Note, however, that in the data you may see special letters encoded using ISO Latin 1 in some loan words, e.g., "société", "l'unità" (these words are not included in CELEX and their analyses will not be evaluated).

Turkish

Standard text. All words are lower-cased. The letters specific to the Turkish language are replaced by capital letters of the standard Latin alphabet, e.g., "açıkgörüşlülüğünü" is spelled "aCIkgOrUSlUlUGUnU".

Arabic

All words in Roman script are presented in ***Buckwalter*** transliteration. The Arabic script is utf-8 coding.

## Download data for Competition 1

| Language | Word list | | Text corpus | Sample of gold standard | Random word pairs file |
|---|---|---|---|---|---|
| English | Text | Text gzipped | Text gzipped | Text | Text |
| Finnish | Text | Text gzipped | Text gzipped | Text | Text |
| German | Text | Text gzipped | Text gzipped | Text | Text |
| Turkish | Text | Text gzipped | Text gzipped | Text | Text |
| Arabic vowelized | Text<br>Arabic script | Text gzipped<br>Arabic script gzipped | Text gzipped<br>Arabic script gzipped | Text | Text |
| Arabic non-vowelized | Text<br>Arabic script | Text gzipped<br>Arabic script gzipped | Text gzipped<br>Arabic script gzipped | Text | Text |

Instead of downloading each file separately, you can download the whole package (including all Competition 1,2 and 3 data), either as a tar file: **morphochal09data.tar** (638 MB; unpack using "`tar xf`") or as a zip file: **morphochal09data.zip** (639 MB).

## Download data for Competition 2

Participation in competition 2 does not necessarily require any extra effort by the participants. The organizers will use the analyses provided by the participants for competition 1 in information retrieval experiments. Data from **CLEF** will be used.

However, because the information retrieval evaluation texts are different from the training texts of competition 1, a slightly better IR performance may be obtained, by submitting also the analyses of the words that do not exist in the word lists of competition 1. The joined word lists can be downloaded below.

| Language | Word list | | Text corpus |
|---|---|---|---|
| English | Text | Text gzipped | See the paragraph below |
| Finnish | Text | Text gzipped | See the paragraph below |
| German | Text | Text gzipped | See the paragraph below |

Those participants who wish to use the full text corpora in order to get information about the context in which the different words occur, please contact the organizers for more information how to register to **CLEF** to obtain the full texts. If there are participants who wish to submit morpheme analysis for words in their actual context (competition 2b), they will need to request the full texts, too. If you need the full texts, please contact the organizers for details how to fill in and submit the **CLEF Registration Form** and **CLEF End-User Agreement**. The DL for this registration is 1 May, 2009.

NOTE: If you do not participate in competition 2b and do not need the full texts for to submit the unsupervised morpheme analysis for competition 2, it is enough to just download the data available at this page.

## Download data for Competition 3

In order to participate in competition 3, participant must submit analysis of the words in the **Europarl corpus**. Two languages, Finnish and Germany, are included in this competition. The result file must be in the same format as in competitions 1 and 2. However, several interpretations per word is not recommended, as only one can be applied. If alternatives are given, we will use only the first one. The word lists can be downloaded below.

| Language | Word list | | Text corpus |
|---|---|---|---|
| Finnish | Text | Text gzipped | Corpus archive (45MB) |
| German | Text | Text gzipped | Corpus archive (54MB) |

Warning: The list of words contains many numbers and various special characters, which may cause probelms if not taken into account. You can preprocess the data if needed, but be careful that the words in the result file will be as they were given. *Exception: It is allowed (and recommended) to change comma (,) to uppercase C. This is necessary especially if your algorithm gives alternative analyses.*

You are free to use the data sets from competitions 1 and 2 in addition to the Europarl set to obtain the analyses. Also, you do not need to return an analysis for every word in the Europarl word list. Those that have no analysis will be treated as one with a single morpheme - the word itself. (Note, however, that Europarl has a large number of words not appearing in the other data sets, so it is not recommended to totally discard it.)

Those participants who wish to use the full text corpora, can use the provided corpus files. The gzipped tar archive contains several hundred text files (named such as `ep-98-01-13.txt`). You must return a set of files that is otherwise the same (same number of lines, same order of lines, *including the empty lines*), but words are replaced by their analyses. Both morphemes and words should be separated by a single space. (I.e., there is no need to distinguish word breaks from other morpheme breaks.)

Page maintained by webmaster at cis.hut.fi, last updated Monday, 09-Mar-2009 17:47:17 EET

## Competition 1

***NEW:*** *The evaluation measures of competition 1 are updated for Morpho Challenge 2009. Some bugs related to the handling of alternative analyses are fixed from the scripts, and points are now measured as one per word, not one per word pair. The new evaluation scripts are now available:*

- ***sample_word_pairs_v2.pl***
- ***eval_morphemes_v2.pl***

*However, the results by the old measures will be provided for a comparison, too. The old scripts are found from **Challenge 2008**.*

In Competition 1, for each language, the morpheme analyses proposed by the participants' algorithm will be compared against a linguistic gold standard. Samples of the gold standards used are available for download on the **datasets** page.

Since the task at hand involves unsupervised learning, it cannot be expected that the algorithm comes up with morpheme labels that exactly correspond to the ones designed by linguists. That is, no direct comparison will take place between labels as such (the labels in the proposed analyses vs. labels in the gold standard). What can be expected, however, is that two word forms that contain the same morpheme according to the participants' algorithm also have a morpheme in common according to the gold standard. For instance, in the English gold standard, the words "foot" and "feet" both contain the morpheme "foot_N". It is thus desirable that also the participants' algorithm discovers a morpheme that occurs in both these word forms (be it called "FOOT", "morpheme784", "foot" or something else).

In practice, the evaluation will take place by sampling a large number of word pairs, such that both words in the pair have at least one morpheme in common. As the evaluation measure, we will use *F-measure*, which is the harmonic mean of *Precision* and *Recall*:

*F-measure* = 1/(1/*Precision* + 1/*Recall*).

*Precision* is here calculated as follows: A number of word forms will be randomly sampled from the result file provided by the participants; for each morpheme in these words, another word containing the same morpheme will be chosen from the result file by random (if such a word exists). We thus obtain a number of word pairs such that in each pair at least one morpheme is shared between the words in the pair. These pairs will be compared to the gold standard; a point is given for each word pair that really has a morpheme in common according to the gold standard. The maximum number of points for one sampled word is normalized to one. The total number of points is 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%.

*Recall* is calculated analogously to precision: A number of word forms are randomly sampled from the *gold standard* file; for each morpheme in these words, another word containing the same morpheme will be chosen from the gold standard by random (if such a word exists). The word pairs are then compared to the analyses provided by the participants; a point is given for each sampled word pair that has a morpheme in common also in the analyses proposed by the participants' algorithm. Points per word is normalized to one and the total number of points is divided by the total number of words.

For words that have several alternative analyses, as well as for word pairs that have more than one morpheme in common, normalization of the points is 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.

### Evaluation of a sample (development test set)

You can evaluate your morphological analyses against the available gold standards (separately for each test language). The program to use for this is the Perl script: **eval_morphemes_v2.pl**. The evaluation program is invoked as follows:

```
eval_morphemes_v2.pl [-trace] wordpairsfile_goldstd wordpairsfile_result
goldstdfile resultfile
```

Four files are given as arguments to `eval_morphemes_v2.pl`:

1. *wordpairsfile_goldstd*: this is the **"random word pairs file"** available for download on the datasets page. This file is needed in the calculation of an estimate of the recall of the proposed morpheme analyses.
2. *wordpairsfile_result*: this file has to be generated using another program (see **below**). It is needed in the calculation of a rough estimate of the precision of the proposed morpheme analyses.
3. *goldstdfile*:this is the **sample of the gold standard** available for download on the datasets page. This file contains the correct morpheme analyses for circa 500 words.
4. *resultfile*: this is the **result file** that your algorithm produces, i.e., a list of words and their proposed morpheme analyses.

The `-trace` argument is optional and produces output for every evaluated word separately. Regardless of the status of the trace argument, the evaluation program produces output of the following kind:

```
PART0. Precision: 69.00% (96/139); non-affixes: 81.55% (51/63); affixes: 58.73% (45/76)
PART0. Recall:    25.59% (142/556); non-affixes: 49.78% (105/211); affixes: 10.78% (37/345)
PART0. F-measure: 37.33%; non-affixes: 61.82%; affixes: 18.22%
#
TOTAL. Precision: 69.00%; non-affixes: 81.55%; affixes: 58.73%
TOTAL. Recall:    25.59%; non-affixes: 49.78%; affixes: 10.78%
TOTAL. F-measure: 37.33%; non-affixes: 61.82%; affixes: 18.22%
```

Note that results are displayed for partition 0 (PART0) and for the entire data (TOTAL). The total scores are here the same as the scores of PART0, since there is only one partition. It is, however, possible to split the data into several partitions and compute results for each partition separately. The overall scores are then calculated as the mean over the partitions. Splitting into partitions is a feature reserved for the final evaluation, when we will assess the statistical significance of the differences between the participants' algorithms.

The figures that count in the final evaluation are the first precision, recall, and F-measure values on the TOTAL lines. These values pertain to all morphemes, but there are also separate statistics for morphemes classified as non-affixes vs. affixes. What counts as an affix is a morpheme with a label starting with a plus sign, e.g., "+PL", "+PAST". This naming convention is applied in the gold standard, which means that you do not have to do anything in order to get the non-affixes/affixes statistics right as far as recall is concerned. However, if you want the same kind of information also for precision, your algorithm must have a means of discovering which morphemes are likely affixes and tag these morphemes with an initial plus sign. Note that it is fully up to you whether you do this or not; it will not affect your position in the competition in any way.

**Sampling word pairs for the calculation of an estimate of the precision**

In order to get an estimate of the precision of the algorithm, you need to provide the evaluation script `eval_morphemes_v2.pl` with a file containing word pairs sampled from your result file. Unfortunately, the estimate is likely to be fairly rough. The reason for this is that you do not have the entire gold standard at your disposal. Thus, if you sample pairs of words that are not included in the 500-word gold standard that you can access, it is impossible to know whether the proposed morphemes are correct or not. What you can do, however, is to make sure that each word that goes into a word pair actually does occur in the 500-word gold standard sample. The problem here is that your algorithm might not propose that many common morphemes for the words within this limited set, and thus the estimate will be based on rather few observations.

Anyway, this is how to do it: First, make a list of *relevant* words, that is, words that are present in the gold standard sample available:

```
cut -f1 goldstdfile > relevantwordsfile
```

Then sample word pairs for 100 words selected by random from your results file:

```
sample_word_pairs_v2.pl -refwords relevantwordsfile < resultfile >
wordpairsfile_result
```

The necessary Perl program is **`sample_word_pairs_v2.pl`**. The output file *wordpairsfile_result* is used as input to `eval_morphemes_v2.pl` (see **above**).

## Competition 2

Competition 2 does not *necessarily* require any extra effort by the participants. The organizers will use the analyses provided by the participants in information retrieval experiments. Data from **CLEF** will be used.
However, those participants who wish to submit morpheme analysis for words in their actual context (competition 2b), please contact the organizers for more information how to register to **CLEF** to obtain the full texts.

In the competition 2 (and 2b) the words in the queries and documents will be replaced by the corresponding morpheme

analyses provided by the participants. We will perform the IR evaluation using the state-of-the-art Okapi (BM25) retrieval method (the latest version of the freely available **LEMUR toolkit**. The most common morphemes in each participant's submission will be left out from the index. The size of this stoplist will be proportional to the amount of the text data in each language and the stoplist size will be the same for each participant's submission. The evaluation criterion will be *Uninterpolated Average Precision*. The segmentation with the highest Average Precision will win. The winner is selected separately for competitions 2 and 2b in each language.

## Competition 3

In competition 3, the morpheme analyses proposed by the participants' algorithm will be evaluated in a statistical machine translation (SMT) framework. The translation models will be trained to translate from a morphologically complex source language to English. The words of the source language will be replaced by their morpheme analyses before training. The translations from this morpheme-to-word model will be combined with translations from a standard word-to-word translation model. For all models, we will use a state-of-the-art phrase-based SMT system. Evaluation of the translations will be performed by applying an automatic metric such as BLEU on a held-out test set.

Data is from the **Europarl corpus**. The participants should apply their algorithms to the list of the word forms in the corpus. It is also possible to use the context information of the words by downloading the full corpus. (See **datasets** for details.)