Morpheme Segmentation Gold Standards for Finnish and English

Mathias Creutz
Helsinki University of Technology*
Mathias.Creutz@hut.fi

Krister Lindén
Helsinki University of Technology,*
University of Helsinki†
Krister.Linden@hut.fi
Krister.Linden@helsinki.fi

*Neural Networks Research Centre, Helsinki University of Technology,
P.O.Box 5400, FIN-02015 HUT, Finland

†University of Helsinki, Department of General Linguistics,
P.O.Box 9, FIN-00014 University of Helsinki, Finland

Abstract

This document describes *Hutmegs*, the Helsinki University of Technology Morphological Evaluation Gold Standard package, which contains gold-standard morphological segmentations for 1.4 million Finnish and 120 000 English words. The Gold Standards comprise surface-string, or allomorph, segmentations of word forms, as well as deep-level, or morpheme, segmentations of the words. The segmentations have been produced semi-automatically and are based on existing resources: the two-level morphological analyzer for Finnish (FINTWOL) and the English CELEX database. For some cases where the transition between two morphemes does not appear clear-cut, so called "fuzzy morpheme boundaries" have been marked as an option. The Hutmegs package also contains some evaluation scripts allowing the user to compute the accuracy compared to the Gold Standard of a segmentation produced by some morphology-learning algorithm. The use of Hutmegs is free for academic purposes, but in order to access the gold-standard segmentations, inexpensive licenses must be purchased from Lingsoft Inc. (for Finnish) and the Linguistic Data Consortium (for English).

1 Introduction

With the emergence of large amounts of textual data in several languages the prospects for designing algorithms that are capable of acquiring language in an unsupervised manner from data seem more and more promising. Also due to the large amounts of data available there is an increasing need for minimally supervised natural language processing systems.

However, a crucial point in the development of data-driven algorithms is the evaluation. The lack of gold-standard references makes it difficult for researchers to assess the performance of their algorithms and to compare them to algorithms developed by others. It is our hope that by providing morphological segmentations for large numbers of Finnish and English word forms, together with software for evaluating segmentation accuracy, we can facilitate the benchmarking and comparison of different algorithms.

In the field of unsupervised learning of the morphology of a natural language, segmentation of word forms into morphemes (or morpheme-like units) is commonly considered an important goal. There are a number of data-driven algorithms that work more or less without supervision and induce, from nothing more than raw text, plausible morpheme segmentations for the words occurring in the text, e.g., (Goldsmith, 2001; de Marcken, 1996; Déjean, 1998; Baayen and Schreuder, 2000; Creutz and Lagus, 2002; Creutz, 2003; Creutz and Lagus, 2004). Morphemes have been defined in linguistic theory as the smallest meaningful units of language. A morpheme corresponds to a meaning and it functions as the smallest element in the syntax of the language (Matthews, 1991). Therefore, morphemes can conceivably be very useful from the point of view of artificial language production or understanding and in applications, such as speech recognition (Siivola et al., 2003), machine translation and information retrieval.

In writing systems where word boundaries are not explicitly marked, word or morpheme segmentation is the first necessary step for any natural language processing task dealing with written text. Languages employing such writing systems comprise, e.g., Chinese and Japanese. The need for common standards for segmentation (in addition to part-of-speech tagging and syntactic bracketing) is apparent. Within the Penn Chinese Treebank project (Xue et al., 2004) a 100 000 word corpus of Mandarin Chinese has been segmented into words, tagged with part-of-speech tags and provided with syntactic bracketing.

In Western languages, there are spaces between the words, and word segmentation of written text is trivial. However, large amounts of work has gone into the annotation of corpora, e.g., part-of-speech tagging, morphological analysis and syntactic bracketing. For American English, the Penn Treebank (Marcus et al., 1993) is an example of existing resources.

A more detailed annotation of the morphological structure of words can be found in the CELEX databases of English, Dutch and German (Baayen et al., 1995). Among

other things, the databases provide information on the derivational and compositional structure as well as inflectional paradigms of tens of thousands of word forms.

Corresponding morphological analyses of word forms, though less detailed, can be obtained using software based on the two-level morphology of Koskenniemi (1983). Such TWOL analyzers exist for, e.g., Finnish, the Scandinavian languages (Swedish, Danish, Norwegian), English and German.[1]

What the existing analyzers and databases lack, however, is an explicit morpheme *segmentation* of the *surface* forms of the words. In the case of the annotated Chinese corpus, the text has indeed been segmented into words. In contrast, the morphological analyses provided by CELEX and TWOL consist of base forms of the words together with morphosyntactic tags. The emphasis of the TWOL is on inflectional morphology. Additionally, a set of tags indicating derivational endings exist and the boundaries within compound words are usually marked. CELEX provides a detailed description of inflectional category together with derivational and compositional structure. This information can be seen as a morpheme segmentation of a word, but the morphemes are not indicated as they are realized on the surface, as word segments or allomorphs, but as deep-level morphemes (or base forms), e.g., the English word 'bacteriologist' yields the segmentation 'bacterium+ology+ist'.

We have produced segmentations of the surface forms of both Finnish and English words. We propose the use of these segmentations as a reference, or Gold Standard, that can be freely used for research purposes, e.g., for the evaluation of unsupervised morphology-learning algorithms or as sub-word units for language modelling in automatic speech recognition. Our work consists in processing the output of the Finnish TWOL and the contents of the English CELEX database to produce an alignment between surface, or allomorph, segmentation and deep-level, or morpheme, segmentation, as in the following examples:[2]

```
tieteellisessä     tietee:tiede|N llise:DN-LLINEN ssä:INE
bacteriologist     bacteri:bacterium|N olog:ology|s ist:ist|s
```

The Finnish Gold Standard contains segmentations for 1.4 million distinct word forms (word types). The English Gold Standard contains segmentations for 120 000 word types. The locations of morpheme boundaries in the surface form is not always obvious and the interpretation chosen by us relies on (Hakulinen, 1979) for Finnish and (Quirk et al., 1985) for English.

The segmentation has been performed semi-automatically with the help of rulesets and a number of scripts. The necessary steps for arriving at the final segmentation are described briefly in the following sections, together with information on formats and access of files and license issues. In addition, we introduce so called "fuzzy morpheme boundaries", which can be applied for cases where it is inconvenient to define one exact

[1] Licenses can be obtained from Lingsoft, Inc. URL: `http://www.lingsoft.fi`

[2] The Finnish word 'tieteellisessä' means 'in [the] scientific'.

transition point between two morphemes.

1.1 Concatenative morphology

A segmentation into morphemes implies that words are formed mainly through a concatenation process. Our Gold Standard relies on the *Item and Arrangement* model of morphology: Words consist of sequences of morphemes and due to some restrictions in the language only a certain realization of a morpheme is possible in a particular context. This realization variant is called an allomorph. An example of allomorphs in Finnish is the inessive case ending, which is realized as '-ssa' or '-ssä' depending on the preceding word stem, e.g., 'talo+ssa', 'metsä+ssä'.

Another type of concatenative morphology is represented by the *Item and Process* model. Also in this model, words are thought to consist of sequences of morphemes. When the morphemes are joined together they trigger morpho-phonological processes, which alter their realization. As an example we can say that the basic form of the Finnish inessive ending is '-ssA', where the capital letter 'A' represents an open unrounded vowel, which becomes the open, unrounded *back* vowel 'a', when the stem contains back vowels; and the open, unrounded *front* vowel 'ä' otherwise.

In our gold-standard segmentation no morpho-phonological processes have been marked explicitly. However, we believe that the Gold Standard can be useful also in the evaluation of algorithms relying on the Item and Process model. For instance, such an algorithm would first segment the words 'talossa' and 'metsässä' into 'talo+ssa' and 'metsä+ssä', and then learn phonological processes, enabling it to conclude that '-ssa' and '-ssä' are realizations of the same morpheme. Since the Gold Standard contains an alignment between allomorphs and their underlying morphemes, it is straight-forward to evaluate how accurately the algorithm discovers these relationships.

As for other models of morphology, which do not assume that words are formed by concatenation, our Gold Standard is less informative. For instance, the patterns of vowel change in so called strong verbs in English have not been indicated, e.g., 'sing'–'sang'–'sung', 'ring'–'rang'–'rung'. All three forms of either verb are merely marked as allomorphs of the base forms 'sing' and 'ring', respectively.

1.2 Fuzzy locations of morpheme boundaries

In some cases, the "linguistically correct" location of a morpheme boundary may not seem the only plausible solution. Historic development of the language may affect the way linguists describe the contemporary morphology. However, from the point of view of natural language applications, this may not be the optimal description.

In our reference segmentations, there is a notation for marking "fuzziness" of morpheme boundaries. The fuzziness consists in alternative locations for the same morpheme boundary, i.e., the boundary does not have an unambiguous location. Users of

our Gold Standard can choose whether they want to evaluate their own segmentation against the one correct "linguistic" segmentation, or against the fuzzy segmentation.

We allow fuzziness as follows: If at the end of a morpheme, there is one phoneme (or sometimes more) that may be totally absent in some allomorphs of the morpheme, this phoneme is considered to lie on a fuzzy boundary between two morphemes. (The latter morpheme is always a suffix.) The phoneme is on the fuzzy boundary only if it alternates phonologically with a "null phoneme", not if it is replaced by another phoneme. This is a somewhat arbitrary definition, but our motivation is that the phoneme (or phonemes) seems to be a seam, or a joint, which is not always needed. If the "joint phoneme" is present only in combination with some following suffixes, it could be considered part of the suffix as easily as part of the preceding morpheme.

For instance, in English, the stem-final 'e' in verbs is dropped in some forms. The user of the Gold Standard can choose whether to consider only the traditional linguistic segmentation correct, as in:

`invite`, `invite+s`, `invit+ed` and `invit+ing`,

or whether also to allow for an alternative interpretation, where the 'e' is considered part of the suffix, as in:

`invit+e`, `invit+es`, `invit+ed` and `invit+ing`,

In the former case, there are two allomorphs of the stem ('invite' and 'invit'), and one allomorph for the suffixes. In the latter case, there is only one allomorph of the stem ('invit'), whereas there are two allomorphs of the third person in the present tense ('-s' and '-es') and an additional infinitive ending ('-e'). Since there are a much greater number of different stems than suffixes in the English language, the latter interpretation lends itself to more compact Item and Arrangement models of morphology. The same reasoning applies to the Finnish language and some examples of fuzzy morpheme boundaries for Finnish will be presented in Section 2.6.

2 Finnish morpheme segmentations

This section describes the steps in the process of acquiring morpheme segmentations for 1.4 million Finnish word forms, as well as the notation used for the final segmentations.

2.1 Data

A 32 million word corpus was compiled containing Finnish text from books and newspapers as well as newswires. The newswires originate from the Finnish National News Agency and the rest of the material from the Finnish Center for Scientific Computation (CSC).[3]

A word list was produced containing each word type occurring in this corpus, i.e., one occurrence of every distinct word form. All upper-case letters were converted to lower-case. Compound words containing hyphens were stored both as entire words and as separate words split at the hyphens. (The latter approach was due to the fact that in Finnish hyphens are almost as obvious delimiters as spaces, and that morphology-learning algorithms may benefit from this knowledge and treat hyphens as word breaks in the pre-processing.) The resulting word list contained 1.7 million word types.

2.2 TWOL analyses

The obtained word list was processed using the FINTWOL analyzer (a morphological transducer lexicon description for Finnish), licensed from Lingsoft, Inc. Out of the 1.7 million word forms, 1.4 million were recognized by TWOL, and for these word forms a morphological analysis was obtained.

The analysis consists of the base form of the word together with morphosyntactic tags. Also tags corresponding to derivational endings exist and the boundaries within compound words are marked with a number sign '#'. In case of ambiguous readings, analyses for all alternative interpretations are produced.

We shall follow the development of the segmentation of a particular word: 'kahvinjuojia' ([some] coffee-drinkers). The TWOL analysis looks like this:

```
"kahvin#juoja" DV-JA N PTV PL
```

The base form is 'kahvinjuoja' (coffee-drinker), which contains an agentive suffix (DV-JA). The word is a noun (N) and inflected into the partitive case (PTV) plural (PL).

2.3 Deep-level morpheme segmentation

The TWOL analysis is no morpheme segmentation. First of all, the base form is a compound, and only its last part ('juojia') has so far been analyzed. The morphological

[3] URL: `http://www.csc.fi/index.phtml.en`

structure of the first part (‘kahvin’) is now obtained by running it alone through the TWOL analyzer, which yields:

```
"kahvi" N GEN SG
```

The base form is ‘kahvi’ (coffee), which is a noun (N) and inflected into the genitive case (GEN) singular (SG).

By appending the two analyzes, we almost obtain a deep-level morpheme segmentation:

```
kahvi N GEN SG juoja DV-JA N PTV PL
```

Next, tags corresponding to no surface morphemes, e.g., the null morpheme for singular (SG), are removed and the part-of-speech information is appended to the preceding stem label (if the preceding label corresponds to a stem). The tags are re-ordered to reflect the order in which the morphemes are realized in the surface form:

```
kahvi|N GEN juoja DV-JA PL PTV
```

This representation now corresponds to a deep-level morpheme representation, or a “base-form morpheme segmentation”. Each morpheme label corresponds to an allomorph that is present in the word form. One problem remains, however: Unlike the TWOL, we want to treat inflection and derivation equally. Thus, the stem morpheme for ‘juojia’ (drinkers) should be ‘juoda’ (to drink) instead of ‘juoja’ (drinker). We will come back to this in Section 2.5 below.

2.4 Surface-level allomorph segmentation

The surface form of the word can be aligned with the deep-level morpheme segmentation in order to get a surface-level segmentation. Each segment will represent an allomorph (a realization variant) of an underlying morpheme. The alignment of our example word ‘kahvinjuojia’ looks like this:

Surface allomorphs	`kahvi`	`n`	`juo`	`j`	`i`	`a`
Underlying morphemes	`kahvi|N`	`GEN`	`juoja`	`DV-JA`	`PL`	`PTV`

In order to obtain a linguistically correct alignment a ruleset has been devised. The ruleset is based on (Hakulinen, 1979) and defines, e.g., all possible allmorph realizations of the suffix tags. The rules state, for instance, that the plural of nominals (PL) is realized as ‘-i-’, ‘-j-’, or ‘-t’.

2.5 Deep-level morpheme segmentation revisited

The morpheme labels preceding derivational endings are not yet correct (cf. Section 2.3 above). At this stage, the surface segmentation is utilized. All instances in all segmented words of the allmorph ‘juo’, except those preceding derivational suffixes, are investigated and the corresponding underlying morphemes are collected. If

the morpheme is unambiguous, such as in the case of ‘juo’, which is always an allomorph of the verb ‘juoda’ (to drink), the labels are completed automatically:

```
kahvi|N GEN juoda|V DV-JA PL PTV
```

In case of ambiguities, such as the allomorph ‘tunne’, which can correspond to either the noun ‘tunne’ (feeling) or the noun ‘tunti’ (hour), the completions have been made manually.

2.6 Fuzzy morpheme boundaries

The motivations for so called “fuzzy morpheme boundaries” have been given in Section 1.2. There are a number of cases where fuzzy boundaries have been allowed as alternative segmentations for Finnish words. In the following, some examples are given:

The proper name ‘Windsor’ has three allomorphs in Finnish: ‘Windsor’ (nominative singular, genitive plural), ‘Windsori’ (oblique cases in singular, nominative plural), and ‘Windsore’ (oblique cases in plural). The following segmentations are linguistically conventional, e.g., ‘Windsor’, ‘Windsori+n’, ‘Windsori+lla’, ‘Windsori+t’, ‘Windsor+i+en’, ‘Windsore+i+lla’. Since the final vowel of the stem is not always present, it belongs to a fuzzy boundary, and can therefore also be attached to the ending: ‘Windsor’, ‘Windsor+in’, ‘Windsor+illa’, ‘Windsor+it’, ‘Windsor+i+en’, ‘Windsor+ei+lla’.

The adjective ‘hapan’ (sour) has four allomorphs: ‘hapan’, ‘happama’, ‘happame’, and ‘happam’. Only the final vowels ‘a’ and ‘e’ in ‘happama’/‘happame’ are on a fuzzy boundary. (The consonants ‘n’ and ‘m’ alternate with each other.) E.g., traditionally we have: ‘hapan+ta’, ‘happama+lla’, ‘happame+sti’, ‘happam+i+a’, ‘happam+uus’; due to fuzziness we also allow: ‘happam+alla’, ‘happam+esti’.

Sometimes the baseform of a word ends in a fuzzy phoneme, e.g., the adjective ‘mukava’ (nice, comfortable) having the allomorphs ‘mukava’ and ‘mukav’. Some traditional segmentations are ‘mukava’, ‘mukava+n’, ‘mukav+i+a’, ‘mukav+uus’. As the final ‘a’ is fuzzy, we obtain the alternative segmentation ‘mukav+an’ for ‘mukavan’. In order to obtain a consistent inflection paradigm, we also allow the segmentation ‘mukav+a’ for ‘mukava’. That is, the baseform is allowed to have an ending ‘-a’, even though there is no ending according to the traditional view.

An extensive list of the cases were fuzzy boundaries are applied for Finnish is shown Table 1.

Table 1: List of all morpheme-final Finnish phonemes that can lie on a fuzzy boundary, and thus are allowed to be attached to either the morpheme preceding or following the morpheme boundary. Examples are given of words segmented into morphemes, where the fuzzy phonemes are present (column labelled "Allomorph with fuzzy phonemes"), and missing (column labelled "Without"). The morphemes in question are rendered in **bold-face** and their part of speech is indicated in the second column from the left. The phonemes on the fuzzy boundary are rendered in *italics*. The selected words illustrate how fuzziness is applied in different inflection paradigms.

Fuzzy phonemes	Part of speech	Example morpheme segmentations	
		Allomorph with fuzzy phonemes	Without
-a	Adj.	**hausk*a***+n, **hausk**+*a*n	**hausk**+uus
	Noun	**koir*a***, **koir**+*a*	**koir**+i+a
	Verb	**laul*a***+vat, **laul**+*a*vat	**laul**+ele+vat
	Suffix	sano+**tt*a***+isi+in, sano+**tt**+*a*isi+in	sano+**tt**+i+in
	Noun	**Tuoma*a***+n, **Tuoma**+*a*n	**Tuomas**
-e	Adj.	**hausk*e***+mpi, **hausk**+*e*mpi	**hausk**+uus
	Noun	**hyttys*e***+stä, **hyttys**+*e*stä	**hyttys**+i+stä
	Noun	**kahv*e***+i+ssa, **kahv**+*e*i+ssa	**kahv**+i+en
	Verb	**päät*e***+tä+än, **päät**+*e*tä+än	**päät**+i+t
	Suffix	laul+**el*e***+vat, laul+**el**+*e*vat	laul+**el**+i+vat
	Noun	**Tuomaks*e***+n, **Tuomaks**+*e*n	**Tuomas**
	Noun	**tähd*e***+n, **tähd**+*e*n	**tähd**+i+stä
	Noun	**venee**+seen, **vene**+*e*seen	**vene**
	Noun	**Windsor*e***+i+lla, **Windsor**+*e*i+lla	**Windsor**
-i	Adj.	**kalli*i***+ksi, **kalli**+*i*ksi	**kallis**
	Noun	**kahv*i***, **kahv**+*i*	**kahv**+i+en
	Noun	**täht*i***, **täht**+*i*	**tähd**+i+stä
	Noun	**Windsor*i***+lla, **Windsor**+*i*lla	**Windsor**
-n	Verb	**vasta*n***+nut, **vasta**+*n*nut	**vasta**+us
-o	Adj.	**hausk*o***+i+ssa, **hausk**+*o*i+ssa	**hausk**+uus
	Verb.	**laul*o***+i+vat, **laul**+*o*i+vat	**laul**+el+i+vat
	Suffix	valitsi+**j*o***+i+ta, valitsi+**j**+*o*i+ta	kannatta+**j**+i+a
-s	Verb	**vasta*s***+i, **vasta**+*s*i	**vasta**+us
-t	Verb	**vasta*t***+a, **vasta**+*t*a	**vasta**+us
	Noun	**vene*t***+tä, **vene**+*t*tä	**vene**
-ä	Noun	**pyör*ä***+ni, **pyör**+*ä*ni	**pyör**+i+ltä
	Adj.	**iäkkä*ä***+llä, **iäkkä**+*ä*llä	**iäkäs**
	Adj.	**jämäkk*ä***+nä, **jämäkk**+*ä*nä	**jämäkk**+yys
	Suffix	jämäkä+**mp*ä***+nä, jämäkä+**mp**+*ä*nä	jämäkä+**mp**+i+in

Table 1: (continued)

Fuzzy phonemes	Part of speech	Example morpheme segmentations Allomorph with fuzzy phonemes	Without
	Verb	**päät*ä***+mme, **päät**+*ä*mme	**päät**+i+mme
-ö	Adj.	**jämäk*ö***+i+tä, **jämäk**+*ö*i+tä	**jämäkk**+yys
-an	Num.	**kahdeks*an***, **kahdeks*a***+*n*, **kahdeks**+*an*	**kahdeks**+i+ssa
-en	Num.	**kymmen*en***, **kymmene**+*n*, **kymmen**+*en*	**kymmen**+i+ä
-ne	Verb	**pahe*ne***+e, **pahe*n***+*e*e, **pahe**+*ne*e	**pahe**+nta+a
-se	Verb	**naris*e***+e, **naris**+*e*e, **nari**+*se*e	**nari**+na
-si	Verb	**naris*i***+ja, **naris**+*i*ja, **nari**+*si*ja	**nari**+na
-ts	Verb	**vali*ts***+i, **vali*t***+*s*i, **vali**+*ts*i	**vali**+nta
-än	Num.	**yhdeks*än***, **yhdeks*ä***+*n*, **yhdeks**+*än*	**yhdeks**+i+ssä
-tse	Verb	**vali*tse***+n, **vali*ts***+*e*n, **vali*t***+*se*n, **vali**+*tse*n	**vali**+nta
-tsi	Verb	**vali*tsi***+ja, **vali*ts***+*i*ja, **vali*t***+*si*ja, **vali**+*tsi*ja	**vali**+nta

2.7 Syntax of the Finnish gold-standard segmentation file

In the Finnish gold-standard segmentation file each row has the format:

```
<wordform><TAB><segmentations><NEWLINE>
```

`<wordform>` and `<segmentations>` can contain any printable characters except space, tab, newline, and carriage return. There are some characters with a special function. If the special function is not intended, the character in question is preceded by the escape character `\` (backslash). Backslash itself is written as `\\` (double backslash). `<wordform>` is the word for which a morphological analysis, i.e., segmentation, is available. All characters are lower-case. `<segmentations>` contains one or several alternative analyses for `<wordform>`. The alternatives are separated using `,` (comma). Each segmentation consists of *chunks*, which are separated by space. Each chunk consists of two parts, separated by `:` (colon). The first part is the *allomorph* (the surface segment of the word) and the second part is the *morpheme*, i.e., the base form or deep-level representation. The format of a segmentation is thus:

```
<allomorph1>:<morpheme1> <allomorph2>:<morpheme2> ...
```

E.g.,

```
arvoamme      arvo:arvo|N a:PTV mme:1PL, arvo:arvo|N amme:amme|N
```

The word ‘arvoamme’ has two interpretations: ‘arvo+a+mme’ (‘of our value’) and the improbable ‘arvo+amme’ (‘valuable bathtub’).

Table 2: Finnish part-of-speech tags.

A	adjective
ADV	adverb
A/N	adjective or noun
CLI	clitic (suffix-like particle, e.g., -kin, -kaan, -pa, -ko)
N	noun
NUM	numeral
PFX	prefix
PRON	pronoun
V	verb

2.7.1 The morpheme part of a chunk

The second part of each chunk in the segmentation can consist of a string in lower-case, which is the base form of the morpheme, usually followed by | (vertical line) and a part-of-speech tag, e.g., `arvo|N`; indicating that 'arvo' (value) is a noun stem. Sometimes the vertical line and part-of-speech tag is missing, which is the case when the part-of-speech information is missing in FINTWOL or when the part of speech does not belong to the major parts of speech listed in Table 2.

The morpheme part of the chunk can also consist of a suffix label, which is written in capital letters, e.g., `PTV`. The first or last character can, however, be a digit: 1, 2, 3, or 4, e.g., `1PL`, and the label can contain a hyphen, e.g., `DV-MA`. Tables 3 and 4 show the possible suffix labels (inflectional suffixes in Table 3 and derivational suffixes in Table 4). Note that for Finnish we have chosen not to include null morphemes, i.e., morphemes that are not realized as segments of words. Therefore there are no suffix labels for, e.g., nominative case, singular number, or present tense.

The morpheme part of a chunk can also consist of a single ∼ (tilde sign), which means that a segment of the word corresponds to *no underlying morpheme*. The only case, where this applies is the hyphen in compounds, such as:

```
viljo-eno     viljo:viljo|N -:∼ eno:eno|N
```

2.7.2 The allomorph part of a chunk

The allomorph part of a chunk can contain either of the special characters ^ (caret) or " (citation mark), which mark fuzzy boundaries. The caret indicates that the following morpheme boundary may come earlier and at the earliest at the point of the caret, e.g.,

```
ilmenevistä     ilme^ne:ilmetä|V v:PCP1 i:PL stä:ELA
```

yields the possible segmentations:

`ilmene+v+i+stä`, `ilmen+ev+i+stä`, and `ilme+nev+i+stä`.

Table 3: The Finnish inflectional suffixes and their labels (most labels are the same as in FINTWOL). The suffixes are grouped into nominal and verbal suffixes. The nominal suffixes are further grouped into case, number, comparison, adverbial, possessive and other suffixes. The case suffixes are additionally grouped into four sub-categories. The verbal suffixes are grouped into tense, voice, mood, infinitive, person and participle suffixes.

INFLECTIONAL SUFFIXES

N o m i n a l

Case

Grammatical		*Location (internal)*		*Location (external)*		*Marginal*	
GEN	genitive	INE	inessive	ADE	adessive	ESS	essive
PTV	partitive	ELA	elative	ABL	ablative	TRA	translative
ACC	accusative (of certain pronouns)	ILL	illative	ALL	allative	INS	instructive
						ABE	abessive
						COM	comitative

Number		Comparison		Adverbial		Other	
PL	plural	CMP	comparative	MAN	manner	PRONSUF	Clitic-like suffix of pronouns e.g., jo+ssa+*kin*
		SUP	superlative	PROL	prolative		

Possessive

1SG	1st person singular	1PL	1st person plural
2SG	2nd person singular	2PL	2nd person plural
3SGPL	3rd person singular or plural		

V e r b a l

Tense		Voice		Mood		Infinitive	
PAST	past tense	PSS	passive	COND	conditional	INF1	1st infinitive
				POTN	potential	INF2	2nd infinitive
				IMPV	imperative	INF3	3rd infinitive
						INF5	5th infinitive

Person		Participle	
SG1	1st person singular	PCP1	present participle
SG2	2nd person singular	PCP2	past participle active
SG3	3rd person singular	PSSPCP2	past participle passive
PL1	1st person plural	DV-MA	agent participle
PL2	2nd person plural		
PL3	3rd person plural		
PE4	4th person (for passive voice)		

Table 4: The labels for the Finnish derivational suffixes (most labels are the same as in FINTWOL). The suffixes are grouped according to the part of speech of the stem that they can be attached to: noun, verb, adjective, or numeral.

DERIVATIONAL SUFFIXES

Noun stem	Verb stem	Adjective stem	Numeral stem	
DN-INEN	DV-AISE	DA-US	ORD	ordinal
DN-ITTAIN	DV-ELE	DA-UUS	FRAC	fractional
DN-LAINEN	DV-ILE			
DN-LAISTA	DV-JA			
DN-LAISTU	DV-MA			
DN-LLINEN	DV-MATON			
DN-MAINEN	DV-MINEN			
DN-TAR	DV-NA			
DN-TON	DV-NEISUUS			
DN-UUS	DV-NTA			
	DV-NTAA			
	DV-NTI			
	DV-SKELE			
	DV-SKENTELE			
	DV-TTA			
	DV-TU			
	DV-U			
	DV-US			
	DV-UTTA			
	DV-UTU			
	DV-VAINEN			

The citation mark indicates that an additional morpheme boundary may be inserted at any point from the location of the citation mark to the end of the morpheme, e.g.,

```
ilmene      ilme"ne:ilmetä|V
```

yields the possible segmentations:

`ilmene, ilmen+e,` and `ilme+ne`.

Thus, as there is no suffix allomorph to which to attach the fuzzy stem-final phonemes, it is necessary to create a new, but optional, suffix. The same applies when the fuzzy boundary is followed by a clitic or a stem. The fuzzy phonemes are not attached to the following clitic or stem, because they clearly do not belong to it. Instead an optional suffix can be inserted, as for:

`kenttävartioon kentt"ä:kenttä|N vartio:vartio|N on:ILL`,
which yields the possible segmentations:
`kenttä+vartio+on` and `kentt+ä+vartio+on`.

3 English morpheme segmentations

This section describes the steps in the process of acquiring morpheme segmentations for almost 120 000 English word forms, as well as the notation used for the final segmentations.

3.1 Data

The English lemma and wordform lexicons in the CELEX database have been used as data for our task. These lexicons contain roughly 70 000 distinct word forms. (The number pertains to the number of word types that are *spelled* differently, without considering the fact that identical word forms may be ambiguous and have several different meanings and morphological structure.)

The CELEX database also contains information on phrasal verbs, e.g., 'shrug off'. These were left out, as we are only concerned with the space-delimited words of English.

3.2 Complete and flat segmentation

A flat, non-hierarchical, segmentation for each word form was extracted from the CELEX database. The segmentation is complete in the sense that it identifies all the morphemes a word form contains. The information gathered for each word contains the morphemes ("deep-level" or baseform) together with part-of-speech and type-of-flection labels. For instance, for the word 'bacteriologists' the following information is obtained:

```
bacterium|N ology|s ist|s N+P
```

That is, the base form of the word consists of three morphemes: 'bacterium', which is a noun (N), and 'ology' and 'ist', which are suffixes (s). The whole word is a noun and inflected into plural (N+P).

3.3 Surface-level allomorph segmentation

The surface form of the word is aligned with the deep-level morpheme segmentation in order to get a surface-level segmentation. Each segment will represent an allomorph (a realization variant) of an underlying morpheme. The alignment of the word 'bacteriologists' looks like this:

Surface allomorphs	`bacteri`	`olog`	`ist`	`s`
Underlying morphemes	`bacterium|N`	`ology|s`	`ist|s`	`N+P`

In order to obtain a linguistically correct alignment for English, we consulted (Quirk et al., 1985).

3.4 Possessive forms of nouns

The English CELEX lexicons do not contain nouns in the possessive form, such as man's, men's, father's, fathers'. In order to provide analyses for word forms in the possessive form, we have automatically generated the possessive forms for all nouns in the data. Nouns in singular ending in 's' receive two variants: a possesive ending only in an apostrophe and a possessive ending in an apostrophe and 's', e.g.,

`Julius'` and `Julius's`.

Nouns in singular and plural ending in another letter than 's' receive the apostrophe followed by 's' as their possessive ending, e.g.,

`man's, father's, mother's` and `men's`.

Nouns in plural ending in 's' receive an apostrophe as their possessive ending, e.g.,

`fathers'` and `mothers'`.

This automatic generation of possessives makes the original word list grow from about 70 000 to almost 120 000 word forms. Many of the resulting words are not likely to occur in real texts, although they do not seem to be incorrect.

3.5 Fuzzy morpheme boundaries

In English, the stem-final 'e' is dropped in some forms. This phenomenon is the only one that fulfills our criterion for "fuzzy boundaries" (cf. Section 1.2).

3.6 Syntax of the English gold-standard segmentation file

The English gold-standard segmentation file follows the same syntax as the Finnish file, but the morpho-syntactic labels are different. Each row in the file has the format:

`<wordform><TAB><segmentations><NEWLINE>`

`<wordform>` and `<segmentations>` can contain any printable characters except space, tab, newline, and carriage return. There are some characters with a special function. If the special function is not intended, the character in question is preceded by the escape character `\` (backslash). Backslash itself is written as `\\` (double backslash). `<wordform>` is the word for which a morphological analysis, i.e., segmentation, is available. `<segmentations>` contains one or several alternative analyses for `<wordform>`. The alternatives are separated using `,` (comma). Each segmentation consists of *chunks*, which are separated by space. Each chunk consists of two parts, separated by `:` (colon). The first part is the *allomorph* (the surface segment of the word) and the second part is the *morpheme*, i.e., the base form or deep-level representation. The format of a segmentation is thus:

```
<allomorph1>:<morpheme1> <allomorph2>:<morpheme2> ...
```

E.g.,

```
dials      dial:dial|N s:N+P, dial:dial|N s:V+e3S
```

The word ‘dials’ has two interpretations: a noun in plural or a present tense verb in the third person singular. (In CELEX the stem ‘dial’ is given as the noun ‘dial’ in both cases with a derivation from noun to verb called conversion. We indicate the conversion with the stem ‘`dial|N`’ having the suffix ‘`s:V+e3S`’. An alternative mark-up could have been ‘`dial|N ∼:V s:e3S`’ indicating an explicit derivational step with an empty surface string.)

3.6.1 The morpheme part of a chunk

The second part of each chunk in the segmentation can consist of a string, which is the base form of the morpheme, followed by | (vertical line) and a part-of-speech tag, e.g., `dial|N`; indicating that ‘dial’ is a noun stem. In some rare cases the part-of-speech tag is missing (when CELEX does not indicate the part of speech). The possible part-of-speech tags are listed in Table 5.

The morpheme part of the chunk can also consist of a label, e.g., `V+e3S`, which indicates both the part of speech of the whole word and the inflection form, which often corresponds to a surface allomorph. The label consists of a part-of-speech tag, which is separated from a string of inflection features by a `+` (plus sign), e.g.,

```
dial:dial|N s:V+e3S.
```

In this interpretation, the whole word ‘dials’ is a verb (`V`), inflected into present tense (`e`), third person (`3`) singular (`S`). Table 6 shows the possible inflection features for English.

The morpheme part of a chunk can also consist of a single ∼ (tilde sign), which means that a segment of the word corresponds to *no underlying morpheme*. The only case, where this applies is the hyphen in compounds, such as:

```
ball-dresses      ball:ball|N -:∼ dress:dress|N es:N+P
```

3.6.2 The allomorph part of a chunk

The allomorph part of a chunk can consist of a single ∼ (tilde sign), which means that the chunk corresponds to a null morpheme, i.e., an inflection feature that is not realized as a segment of the word, e.g.,

```
dress      dress:dress|N ∼:N+S, dress:dress|N ∼:V+i.
```

Here the segmentation of the word ‘dress’ consists of only one segment, which consists of the entire word. The word can be interpreted as a noun in singular (`N+S`) or as the

Table 5: English part-of-speech tags. (The tags are the same as in CELEX.)

A	adjective
B	adverb
C	conjunction
D	article
I	interjection
N	noun
O	pronoun
P	preposition
Q	numeral
V	verb
p	prefix
s	suffix
?	undetermined

Table 6: English inflection features. (All tags except the possessive tag have been adopted directly from CELEX.) The features are grouped according to part of speech.

Nouns		Verbs		Adjectives and adverbs		Other	
S	singular	S	singular	b	positive	X	headword form
P	plural	e	present tense	c	comparative		
o	possessive	a	past tense	s	superlative		
r	rare form	i	infinitive				
		p	participle				
		1	1st person				
		2	2nd person				
		3	3rd person				
		r	rare form				

infinitive of a verb (`V+i`), but neither of these morphological features are realized as a word segment.

The allomorph part of a chunk can contain either of the special characters ^ (caret) or " (citation mark), which mark fuzzy boundaries. The caret indicates that the following morpheme boundary may come earlier, at the point of the caret, e.g.,

```
loves      lov^e:love|V s:V+e3S
```

yields the possible segmentations:

`love+s` and `lov+es`.

The citation mark indicates that an additional morpheme boundary may be inserted at the citation mark, e.g.,

```
love      lov"e:love|V ∼:V+i
```

yields the possible segmentations:

`love` and `lov+e`.

Thus, as there is no suffix allomorph to which to attach the fuzzy stem-final ‘e’, it is necessary to make a suffix of the ‘e’. The same applies when the fuzzy boundary is followed by a morpheme that is not a suffix. The fuzzy ‘e’ is not attached to the following morpheme, because it clearly does not belong to it. Instead an ‘e’-suffix can be inserted, as for:

```
lovebird      lov"e:love|V bird:bird|N ∼:N+S
```

which yields the possible segmentations:

`love+bird` and `lov+e+bird`.

4 Installation and evaluation

This section describes how to obtain the gold-standard segmentations for Finnish and English and install them on your computer. A Linux or Unix operating system is necessary. Instructions are also given regarding a few scripts, which are helpful in the evaluation of a segmentation produced by some algorithm.

4.1 Download and licensing conditions

The *Helsinki University of Technology Morphological Evaluation Gold Standard* package, *Hutmegs*, is a collection of files and documentation that are free to use for non-commercial purposes as long as a reference is made to the source of the material.[4] The Hutmegs package can be downloaded from the following Internet address: `http://www.cis.hut.fi/projects/morpho/`.

However, the Finnish gold-standard segmentations are based on the FINTWOL analyzer, which is a commercial product. To obtain the complete Finnish Gold Standard, a missing component must be licensed from Lingsoft, Inc.[5] in Helsinki, Finland. The missing component is a file containing the FINTWOL analyses for the word forms comprised in the Gold Standard. The list of FINTWOL analyses has the product name 'LS Hutmegs-Fintwol' (version 1.0) and the one-time license fee for non-commercial activities is 600 euros (as of September, 2004). To purchase this component contact Lingsoft through e-mail: `info@lingsoft.fi`. If the component is not purchased, the user will have access to all Hutmegs scripts and documentation, but only a sample Gold Standard containing the analyses of no more than 700 Finnish word forms.

Likewise, the CELEX database (Baayen et al., 1995) is a prerequisite for accessing the complete English Gold Standard. Two files from CELEX describing the English morphology must be available in order to generate the English gold-standard segmentations. Non-commercial licenses are available from the Linguistic Data Consortium[6] in Philadelphia, USA. The product name of the database is CELEX2 and its catalogue number is LDC96L14.[7] The non-member price is 150 US dollars (as of September, 2004). The Hutmegs package provides sample gold-standard segmentations for roughly 600 English word forms, which can be viewed without access to the CELEX database.

[4] Cite this work like this: "Mathias Creutz and Krister Lindén. 2004. *Morpheme Segmentation Gold Standards for Finnish and English*. Technical Report A77, Helsinki University of Technology, October. URL: `http://www.cis.hut.fi/projects/morpho/`"

[5] URL: `http://www.lingsoft.fi`

[6] URL: `http://wave.ldc.upenn.edu/`

[7] URL: `http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14`

4.2 Installation

The downloaded Hutmegs package version 1.0 is unpacked using the following command:

`tar xzf hutmegs_ver1.0.tar.gz`.

A directory tree will be created with the root `hutmegs` and the subdirectory `ver1.0` (or later version). This directory contains the following:

`README`	Some useful information.
`Makefile`	A Makefile showing examples of how to decode the encoded Finnish and English Gold Standards and how to use the evaluation scripts (described below).
`bin/`	A directory containing all installation and evaluation scripts described in this document. The scripts are *Perl* scripts.
`lib/`	A directory containing the encoded Finnish and English Gold Standards as well as the decoded samples. This directory also contains example morpheme category files for Finnish and English, which can be used in connection with the `boundary_detection_stats.pl` and `evaluate_tags.pl` evaluation scripts.
`test/`	A directory containing example files utilized by the Makefile.
`doc/`	A directory containing this document.

4.2.1 Configuring the scripts

The scripts in the `bin/` directory are Perl scripts. The first line in every script needs to point to the location of the Perl interpreter on the computer system. The value is by default set to `/usr/bin/perl`. If this is incorrect, either edit the scripts in the `bin/` directory manually, or adjust the value of the `PERLCMD` variable in `Makefile`, and run '`make perlconfig`'. This command will update the scripts automatically with the value of the `PERLCMD` variable.

4.2.2 Decoding the Finnish Gold Standard

Once the FINTWOL output file has been purchased from Lingsoft, Inc., the decoding of the Finnish Gold Standard is invoked from the command line using the following syntax:

`decodefintwol.pl fintwol_output encodedgoldstd.fin > goldstd.fin`.

Make sure that the FINTWOL output file `fintwol_output` is unpacked. The `decodefintwol.pl` script is in the `bin/` directory and the `encodedgoldstd.fin` file is in the `lib/` directory.

4.2.3 Decoding the English Gold Standard

The compilation of the English Gold Standard involves two steps. First, run:

`decodecelex.pl eml.cd emw.cd encodedgoldstd.eng > goldstd.tmp`,

where `eml.cd` and `emw.cd` are files from the English CELEX database. The locations on the CD-ROM of these files are `english/eml/eml.cd` and `english/emw/emw.cd`, respectively. The `encodedgoldstd.eng` file is in the Hutmegs `lib/` directory.

Next, run:

`postprocesscelex.pl < goldstd.tmp > goldstd.eng`.

The scripts `decodecelex.pl` and `postprocesscelex.pl` are in the Hutmegs `bin/` directory.

4.3 Evaluation scripts

The four evaluation scripts are found in the Hutmegs `bin/` directory. The scripts are `align_segmentations.pl`, `boundary_precision_recall.pl`, `boundary_detection_stats.pl`, and `evaluate_tags.pl`.

4.3.1 `align_segmentations.pl`

The script `align_segmentations.pl` produces an optimal morpheme-boundary alignment between the gold-standard segmentation of words and the segmentations produced by some algorithm. This alignment is needed as input to the other scripts described below. Invoke `align_segmentations.pl` from the command line using the following syntax:

`align_segmentations.pl [-fuzzy] goldstdfile < segs_2b_evaluated`.

The `-fuzzy` switch is optional and activates the use of fuzzy morpheme boundaries. That is, for some morpheme boundaries several locations are considered correct. When the `-fuzzy` switch is not utilized, all morpheme boundaries have an unambiguous, linguistically conventional, location (cf. Section 1.2). The parameter `goldstdfile` indicates the location of the appropriate gold-standard file (Finnish or English), and `segs_2b_evaluated` indicates the location of the file containing the segmentations to be evaluated.

Input

Each line of `segs_2b_evaluated` must comply with the following format:

`<segmentation><TAB><wcount><NEWLINE>`,

where the segmentation consists of chunks separated by space, and each chunk consists of two parts separated by `:` (colon). The first part is a segment of the word and the second part is some tag, a class assigned by the algorithm to the segment. This corresponds to the allomorph and morpheme parts in the chunks in the gold-standard segmentations (cf. Sections 2.7 and 3.6). The word count (`wcount`) is a positive integer indicating the number of times the word form has occurred in the data processed by the algorithm.

For instance, here is an extract of the segmentation file produced by the so called Category-Learning algorithm presented in (Creutz and Lagus, 2004). The data used consisted of a subset of the Brown corpus comprising 250 000 word tokens:

```
aspect:STM     13
aspect:STM s:SUF       17
asphalt:STM    2
aspir:STM ant:SUF      1
aspir:STM ation:SUF    2
aspir:STM ation:SUF s:SUF      2
aspir:STM ed:SUF       1
aspir:STM es:SUF       1
aspir:STM ing:SUF      1
```

The Category-Learning algorithm works in an unsupervised way and proposes a morpheme segmentation for every word form in the data. Additionally, each morpheme is tagged with one of the categories prefix (PRE), stem (STM), and suffix (SUF). In this example the word ‘aspires’, which has occurred once in the data, has been segmented into ‘aspir+es’. The first segment has been tagged as a stem and the second segment as a suffix.

Output

Each row in the output of `align_segmentations.pl` has the following format:

`<word><TAB><goldstd_seg><TAB><algorithm_seg><TAB><wcount><NEWLINE>`.

The first field contains the word itself. The second field contains the gold-standard segmentation of the word aligned against the proposed segmentation in the third field. The fourth field indicates the word count, i.e., the number of times the word occurred in the data. When the gold-standard segmentation is aligned against the segmentation proposed by the algorithm, there will always be as many chunks in both representations. If necessary, either or both segmentations will be extended with “null chunks” ($\sim$:$\sim$).

For instance, the following output is produced for the word ‘aspires’, when the `-fuzzy` switch is *not* in use:

```
aspires    ~:~ aspire:aspire|V s:V+e3S    aspir:STM ~:~ es:SUF    1
```

The alignment can be presented more explicitly as:

Gold Standard:	∼:∼	aspire:aspire|V	s:V+e3S
Algorithm:	aspir:STM	∼:∼	es:SUF

Null chunks on the gold-standard side can be interpreted as *insertions* of incorrect morpheme boundaries. In this case, the algorithm has proposed an incorrect boundary at the end of 'aspir'. Null chunks in the segmentation proposed by the algorithm can be interpreted as *deletions* of desired boundaries. In the example, the algorithm has missed the boundary at the end of 'aspire'. (Note that the alignment is an optimal alignment of morpheme *boundaries* and not an optimal alignment of morphemes. In our implementation, two morphemes must end at the same point in order to be aligned with each other, which is the case for 's' vs. 'es' above, but which is not the case for 'aspire' vs. 'aspir'.)

In this particular example, the alignment is different, if the `-fuzzy` switch is activated. The stem-final 'e' can then be attached to the suffix, which makes the segmentation proposed by the algorithm correct. Consequently, there are no insertions or deletions:

```
aspires   aspir:aspire|V es:V+e3S   aspir:STM es:SUF   1
```

When there are many alternative correct segmentations in the Gold Standard, there may be several equally good alignments with the single segmentation proposed by the segmentation algorithm. This is represented in the output as several gold-standard segmentations separated by , (comma) in the `<goldstd_seg>` field. These are matched with equally many alignments of the proposed segmentation in the `<algorithm_seg>` field. In the following example the algorithm has incorrectly segmented the word 'conjunction' into two stems 'conj'+'unction'. The correct segmentations, according to the Gold Standard, are 'con+junction' (prefix + noun stem) and 'conjunct+ion' (verb stem + suffix). The long line has been broken onto three lines after the first and second field:

```
conjunction   ...
con:con|p ∼:∼ junction:junction|N, ∼:∼ conjunct:conjoin|V ion:ion|s...
∼:∼ conj:STM unction:STM, conj:STM ∼:∼ unction:STM   2
```

4.3.2 `boundary_precision_recall.pl`

The script `boundary_precision_recall.pl` computes accuracy statistics for the morpheme boundaries proposed by the algorithm. First, an alignment against the gold-standard segmentation must be produced using the script `align_segmentations.pl`. Then, `boundary_precision_recall.pl` is invoked from the command line as follows:

```
boundary_precision_recall.pl < alignment,
```

where `alignment` is the output of `align_segmentations.pl`. The script `bound-`

`ary_precision_recall.pl` only evaluates the placing of the morpheme boundaries, i.e., the segmentation into allomorphs of the surface form of words. The underlying morphemes and their labels play no role whatsoever in this evaluation.

The output of the script looks like the following (explanations are given after the example):

```
Number of word tokens: 250000, of which omitted 9070 (3.63%)
Token precision: 70.36%
Token recall: 64.70%
Token F-measure: 67.41%
Number of word types: 21434, of which omitted 3864 (18.03%)
Type precision: 76.85%
Type recall: 66.13%
Type F-measure: 71.09%
Number of desired morph types: 8749; recognized: 8999
All recognized morph types: 11604, of which omitted 2605 (22.45%)
```

The evaluation statistics are computed both on word *tokens* and word *types*. *Precision* is the proportion of morpheme boundaries suggested by the algorithm that are correct. *Recall* is the proportion of morpheme boundaries in the Gold Standard that were correctly recognized by the algorithm. The *F-measure* is the harmonic mean of precision and recall [see e.g., (Manning and Schütze, 1999)]:

$$\textit{F-Measure} = 1/[\frac{1}{2}(\frac{1}{\textit{Precision}} + \frac{1}{\textit{Recall}})].$$

In the evaluation statistics for word tokens, the segmentation of each occurrence of every word in the data is taken into account. That is, words occurring many times have a higher impact on the result than rare words. In the evaluation statistics for word types, only one occurrence of each distinct word form is considered. That is, the accuracy of the segmentation of rare words dominate, since most of the words are rare.

The number and proportion of *omitted word tokens and types* correspond to words in the data, for which there was no gold-standard segmentation available. These words are left out of the evaluation.

The number of *desired morph types* indicates how many distinct allomorphs (segments) there are in the gold-standard segmentations of the words in the data set used. The number of *recognized* morph types indicates how many distinct allomorphs there are in the segmentations of these words proposed by the algorithm. If these two figures are roughly equal, the algorithm has managed to segment the words to the same degree as in the Gold Standard. That is, on the average, the words have neither been split excessively into many short segments, nor too reluctantly into too long and few segments per word.

The number of *all recognized morph types* indicates how many distinct allomorphs there are in the segmentations proposed by the algorithm for all words in the data, including those words that have been omitted in the evaluation, because there is no gold-standard segmentation available for them.

4.3.3 `boundary_detection_stats.pl`

The script `boundary_detection_stats.pl` is helpful, when it is necessary to know roughly where morpheme boundaries have been detected correctly and incorrectly. The script is invoked as follows:

`boundary_detection_stats.pl morpheme_categories < alignment`,

where `alignment` is the output of `align_segmentations.pl`, which must be run first. The file `morpheme_categories` contains a grouping of morphemes in the Gold Standard into broader categories. This is one possible grouping for the English morphemes (consult Tables 5 and 6 for an explanation of the tags):

```
# All labels corresponding to surface allomorphs are
# grouped into two classes:  stems and affixes
#
STEM   A B C D I N O P Q V ?  # Part-of-speech tags for stems
AFFX   p s                    # Derivational pre- & suffixes
AFFX   A+c A+s B+c B+s        # Comparatives and superlatives
                              # for adjectives and adverbs
AFFX   N+P N+Po N+So          # Plurals & possessives of nouns
AFFX   V+a1S V+e3S V+pe       # Verb endings (past, pres., -ing)
AFFX   ∼                      # Hyphen or extra ending
                              # due to fuzziness
```

Each line of the file contains two tab-separated fields. The left field contains a morpheme label for a broad morpheme category; in this example `STEM` for stems and `AFFX` for affixes. The right field consists of space-separated labels for part-of-speech or suffix type. The number sign (#) indicates the start of a comment. Everything from the number sign to the end of the line is ignored, including any preceding whitespace (tab or space).

The part-of-speech tags and suffix labels correspond to information in the morpheme parts of the chunks in the Gold Standard. If the morpheme part of the chunk contains a base form as for, e.g., `aspire|V`, only the part-of-speech tag (`V`) is retained. The question mark (`?`) covers the cases, where there is a baseform, but the part-of-speech tag is missing. In the current example, word stems of any part of speech are classified as `STEM`, whereas derivational prefixes and suffixes (e.g., `mis|p` and `ion|s`) are classified as `AFFX`.

If the morpheme-part of the chunk contains no baseform, but only a suffix label, the entire label is retained (e.g., `A+c` representing the comparative ending '-er'). In the example, all such cases are classified as `AFFX`. The special character ∼ (tilde) covers cases, where there is a segment in the surface-form of the word, but no underlying morpheme. Hyphens in compounds are such a case. The other case occurs if fuzzy morpheme boundaries are activated and consists in the optional endings at the end of baseforms of words, e.g., the '-e' of 'aspir+e'.

The script evaluates only the *segmentation* proposed by the algorithm, not the tag-

ging. The output looks like this:

```
Incorrectly inserted boundaries (#insertions/#morphemes)
AFFX    tok: 2288/62154 = 0.0368   typ: 518/14215 = 0.0365
STEM    tok: 15728/244871 = 0.0642 typ: 2462/18319 = 0.1344
--
Recall for detection of beginning of morpheme
AFFX    tok: 39634/57639 = 68.76%  typ: 8859/12923 = 68.56%
STEM    tok: 3132/8456 = 37.04%    typ: 1037/2041 = 50.78%
--
Recall for detection of ending of morpheme
AFFX    tok: 6431/10988 = 58.52%   typ: 1972/3312 = 59.56%
STEM    tok: 36335/55106 = 65.94%  typ: 7923/11652 = 68.00%
--
Recall for detection of morpheme transitions
AFFX AFFX   tok: 4608/6888 = 66.89%    typ: 1371/2088 = 65.65%
AFFX STEM   tok: 1823/4100 = 44.46%    typ: 602/1224 = 49.16%
STEM AFFX   tok: 35026/50750 = 69.02%  typ: 7488/10834 = 69.12%
STEM STEM   tok: 1309/4356 = 30.05%    typ: 435/818 = 53.20%
--
```

Statistics are computed both for word *tokens* (`tok`) and *types* (`typ`). First, the distribution of incorrectly inserted morpheme boundaries is shown. For instance, there were a total number of 62 154 morpheme tokens classified as `AFFX` according to the Gold Standard and the morpheme category file. The algorithm proposed 2288 incorrect boundaries within the affix morphemes, resulting in an average of 0.0368 insertions per affix. For the stem morphemes the proportion of insertions was higher, 0.0642, which is natural, since the stems are typically longer than the affixes, and there is thus more "room" for insertions.

Next, the recall for the detection of the beginning of morphemes is shown. (The first morpheme in every word is left out from this calculation, since the beginning of the first morpheme is trivial to detect.) For instance, there were 2041 non-word-initial stem morphemes in the list of word types. The algorithm managed to place a morpheme boundary at the beginning of these morphemes in 1037 cases, representing 50.78 % of all the cases.

Likewise, the recall for the detection of the ending of morphemes is computed. (Here the last morpheme of every word is left out of the calculation.) For instance, there were 11 652 non-word-final stem morphemes in the list of word types. The algorithm detected correctly 7923, or 68.00 %, of the end of these morphemes.

Finally, the placing of boundaries between morphemes is evaluated. According to the example, the most "easy boundaries" to detect are those between a stem and a following affix; 69.02 % were detected (when looking at word tokens). The most difficult transitions to detect were those between two adjacent stems: only 30.05 % were found (word tokens).

4.3.4 `evaluate_tags.pl`

The script `evaluate_tags.pl` evaluates the tagging proposed by the algorithm, and is invoked as follows:

`evaluate_tags.pl morpheme_categories < alignment`.

`alignment` is the output of `align_segmentations.pl`, which must be run first. The file `morpheme_categories` contains a grouping of morphemes in the Gold Standard into the categories learned by the algorithm. Our example algorithm tags the segments it discovers as prefixes, stems, or suffixes (`PRE`, `STM`, or `SUF`). It is thus necessary to map the category labels used in the Gold Standard file onto these three categories. Exactly the same syntax applies as for the `morpheme_categories` file in Section 4.3.3:

```
PRE   p                      # (Derivational) prefix
STM   A B C D I N O P Q V ?  # Part-of-speech tags for stems
SUF   s                      # Derivational suffix
SUF   A+c A+s B+c B+s        # Comparatives and superlatives
                             # for adjectives and adverbs
SUF   N+P N+Po N+So          # Plurals & possessives of nouns
SUF   V+a1S V+e3S V+pe       # Verb endings (past, pres., -ing)
SUF   ~                      # Hyphen or extra ending
                             # due to fuzziness
```

The output of the script looks like this (explanations are given after the example):

```
Number of word tokens: 250000, of which omitted 9070 (3.63%)
Number of word types: 21434, of which omitted 3864 (18.03%)
Token #tags(PRE)    desi: 4537 (1.48%)     reco: 3092 (1.02%)
Token #tags(STM)    desi: 244871 (79.76%)  reco: 246071 (81.56%)
Token #tags(SUF)    desi: 57616 (18.77%)   reco: 52548 (17.42%)
Token #tags(All)    desi: 307025 (100.00%) reco: 301712 (100.00%)
Type  #tags(PRE)    desi: 1304 (4.01%)     reco: 823 (2.70%)
Type  #tags(STM)    desi: 18319 (56.31%)   reco: 19237 (63.18%)
Type  #tags(SUF)    desi: 12910 (39.68%)   reco: 10387 (34.12%)
Type  #tags(All)    desi: 32533 (100.00%)  reco: 30446 (100.00%)
Number of correctly segmented word tokens: 206884 (85.87%)
Number of correctly segmented word types: 11425 (65.03%)
Token tags correct: 240764/242003 (99.49%)
Type tags correct: 18629/18963 (98.24%)
Token precision (PRE): 96.39% (909/943)
Token recall (PRE): 76.90% (909/1182)
Token F-measure (PRE): 85.55%
Token precision (STM): 99.45% (207912/209061)
Token recall (STM): 99.96% (207912/208001)
Token F-measure (STM): 99.70%
Token precision (SUF): 99.83% (31943/31999)
Token recall (SUF): 97.33% (31943/32819)
Token F-measure (SUF): 98.56%
```

```
Type precision (PRE): 96.41% (295/306)
Type recall (PRE): 76.42% (295/386)
Type F-measure (PRE): 85.26%
Type precision (STM): 97.41% (11792/12106)
Type recall (STM): 99.83% (11792/11811)
Type F-measure (STM): 98.60%
Type precision (SUF): 99.87% (6542/6551)
Type recall (SUF): 96.70% (6542/6765)
Type F-measure (SUF): 98.26%
```

The figures are as usual calculated for both word *tokens* and word *types*. First, the number of word tokens and word types in the data are shown together with the number and proportion of word forms for which there is no gold-standard segmentation, and consequently are left out of the evaluation. Then the number and proportion of morphemes tagged with each of the three tags (`PRE`, `STM`, `SUF`) are shown for both the gold-standard segmentation (in the "`desi`red" field) and the segmentation proposed by algorithm (in the "`reco`gnized" field). Ideally, the figures in both fields should match.

The evaluation that follows after this is based *only* on the words that have been *segmented correctly*. The number and proportion of these correctly segmented word tokens and types are shown: 85.87 % (tokens), 65.03 % (types). For the correctly segmented words, 99.49 % of the tags proposed by the algorithm match the tags in the Gold Standard (word tokens). Evaluated on word types, the tagging is correct for 98.24 % of the morphemes.

Finally, precision, recall and F-measure is calculated for each morpheme category separately, on word tokens as well as word types. For instance, the precision for prefixes is 96.39 % (word tokens), which means that 96.39 % of the morphemes tagged as prefixes by the algorithm are indeed prefixes (when only the correctly segmented words are considered). Recall for prefixes is lower, 76.90 % (word tokens), indicating that there were a number of prefixes in the Gold Standard that were tagged as something else by the algorithm.

5 Discussion

The gold-standard segmentations have been produced semi-automatically. The quality of the FINTWOL analyzer as well as the quality of the English CELEX database determine to a large degree the quality of the final gold-standard segmentations. FINTWOL has some tendencies to over-generate, especially for compound words. This means that there can be many alternative analyses for a word, some of which may be peculiar and improbable. Also the stem-final fuzziness applied to the Finnish words has been produced using automatic rules, which made it impossible to consider every case separately due to time constraints.

Derivational morphology is more exhaustively described for English in CELEX than for Finnish in FINTWOL. Finnish derivational morphemes have been provided in the Gold Standard, when the information is available in FINTWOL. Additionally, some obvious mistakes have been corrected for Finnish.

We use the CELEX data to identify the exact morpheme segmentation in the surface string of words. For instance, for the word 'conductivities' we not only satisfy ourselves with the fact that it is derived from 'conduct', but the word is split into the segments 'conduct+iv+iti+es'. Others have used CELEX as well, e.g., (Schone and Jurafsky, 2001; Baayen and Schreuder, 2000), but they do not aim at evaluating a precise segmentation. Schone and Jurafsky (2001) evaluate conflation sets, e.g., the word 'conduct' is derivationally related to 'conduct', 'conducts', 'conducted', 'conducting', 'conduction', 'conductive', 'conductivity', etc. Baayen and Schreuder (2000) use CELEX, but they evaluate manually and any segmentation which is not wrong or improbable is deemed acceptable regardless of the number of segments discovered.

CELEX sometimes gives derivations with morphemes, which have no corresponding surface realization in the final word form. For example, the word form 'conductive' is derived from 'conduct' via the intermediary form 'conduction'. In those cases we stay true to CELEX by introducing a corresponding empty string, or null morpheme, indicated by $\sim$ (tilde) on the surface, e.g., 'conduct+$\sim$+ive'. (The deep-level morpheme segmentation is here: 'conduct|V+ion|s+ive|s'.) However, derivational morphemes corresponding to empty strings are impossible to discover with current segmentation algorithms, so we drop those morphemes before evaluation. If someone invents a method for discovering them, they can easily be retained. At this point, the Finnish and English Gold Standards differ. Whereas there are some null morphemes in the analyses of the English words, we have excluded all null morphemes from the analyses of Finnish words.

Any null morphemes are ignored by the current evaluation scripts (cf. Section 4.3). However, when fuzzy boundaries are allowed, we could make use of these morphemes: For instance, English verbs usually do not have any ending marking the infinitive. But if the final 'e' is detached from the verb stem due to fuzziness, it could be considered as an infinitive ending instead of, as in the current implementation, being treated as

an optional “anonymous” ending (cf. Sections 3.6.2 and 4.3.3). Corresponding cases exist for Finnish: The nominative case of nominals could obtain an ending due to fuzzy boundaries, and so could the present tense of verbs.

There are special cases when the inflectional information is in the middle of the string like ‘aides-de-camp’ as a plural form of ‘aide-de-camp’. These cases are rare in English, so we found it to be most consistent with the Item and Arrangement model not to split them at all, but to treat them as irregular forms, or allomorphs. Compare the inflection of the words ‘sing’ as ‘sing’, ‘sang’, ‘sung’. We do not split this into ‘s+?+ng’, but instead we consider the baseform to have three allomorphs ‘sing’, ‘sang’, ‘sung’, which are selected in their entirety in different contexts.

For other languages, such as Arabic or Semitic languages with frequent infixes, one would likely choose a different segmentation standard. Already the Finnish language exhibits some “infix-like behaviour”, as there are sometimes inflectional “suffixes” in the middle of words. In numeral expressions each number is inflected separately even if the expression is written without spaces in one sequence, e.g., ‘kuude+*lla*+kymmene+*llä*+neljä+*llä* euro+*lla*’ (with sixty-four euros). Some pronouns also have their inflection morphemes in the middle of the word, e.g., ‘jo+*lla*+kin’ (‘with something’), which is the segmentation in our Gold Standard.

To conclude, Hutmegs is an evaluation package for morphological segmentations of Finnish and English vocabularies. It is mainly based on an Item and Arrangement model of morphology, which results in particular strengths and weaknesses. The possibility to use fuzzy morpheme boundaries may alleviate some of the rigidness that would ensue if one single correct answer had to be chosen.

We hope that the growing circle of researchers studying morphology induction will discover the Hutmegs package and find it useful for their work. By providing a segmentation standard that can be used for benchmarking and evaluating different morphology-learning algorithms, we wish to contribute to the promotion of research in this fascinating field.

Acknowledgements

We are grateful to D.Sc. (Tech.) Krista Lagus for her thorough examination of the manuscript and her insightful comments. Furthermore, we would like thank the Graduate School of Language Technology in Finland for funding our work, and we express our gratitude to the Finnish Center for Scientific Computation and the Finnish National News Agency for providing text corpora and making it possible for us to produce a large vocabulary of Finnish words. Our sincere thanks go to Juhani Reiman, CEO of Lingsoft, Inc., for sharing our interest in making the Finnish word segmentations public for research, and to Mr. Sami Virpioja for drawing our attention to a few bugs in earlier versions of the Finnish Gold Standard.

References

R. Harald Baayen and Robert Schreuder. 2000. Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences)*, 358:1–13.

R. Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. URL: `http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14`.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the 6th Meeting of the ACL Special Interest Group of Computaional Phonology (SIGPHON)*, pages 21–30, Philadelphia, Pennsylvania, USA.

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

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

Carl G. de Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, MIT.

Hervé Déjean. 1998. Morphemes as necessary concept for structures discovery from untagged corpora. In *Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299, Adelaide.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Lauri Hakulinen. 1979. *Suomen kielen rakenne ja kehitys (The structure and development of the Finnish language)*. Kustannus-Oy Otava, 4 edition.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Ph.D. thesis, University of Helsinki.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19.

P. H. Matthews. 1991. *Morphology*. Cambridge Textbooks in Linguistics, 2nd edition.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, Essex.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics NAACL-2001*.

Vesa Siivola, Teemu Hirsimäki, Mathias Creutz, and Mikko Kurimo. 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pages 2293–2296, Geneva, Switzerland.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2004. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 10(4):1–30. URL: `http://www.cis.upenn.edu/%7Echinese/`.