To split or not, and if so, where? Theoretical and Empirical Aspects of Unsupervised Morphological Segmentation

Amit Kirschenbaum

Natural Language Processing Group,
Leipzig University, Germany
amit@informatik.uni-leipzig.de

Abstract. The purpose of this paper is twofold: First, it offers an overview of challenges encountered by unsupervised, knowledge free methods when analysing language data (with focus on morphology). Second, it presents a system for unsupervised morphological segmentation comprising two complementary methods that can handle a broad range of morphological processes. The first method collects words which share distributional and form similarity and applies Multiple Sequence Alignment to derive segmentation of these words. The second method then analyses less frequent words utilizing the segmentation results of the first method. The challenges presented in the theoretical part are demonstrated exemplarily on the workings and output of the introduced unsupervised system and accompanied by suggestions how to address them in future works.

1 Introduction

Unsupervised, knowledge free approaches analyse raw, unannotated data without any previous knowledge about the language they are applied on. In the context of morphology, which is at the focus of the present study, this can comprise various tasks like paradigm extraction, detection of related sets of words, or morphological segmentation. In the last half of a century of research, several dozens of algorithms addressing these tasks have been developed (see [15] for an overview), with the result that "high accuracy by ULM systems is presently only achievable if the language has small amounts of one-slot concatenative morphology" [15, p.335].

In this paper, we introduce a two-method system performing morphological segmentation, i.e. splitting word forms of a given language into their basic units carrying meaning - the morphemes. The first method utilizes Multiple Sequence Alignment (MSA). The approach has its origin in bioinformatics, where it is used to align sequences of DNA, RNA, proteins, etc. MSA is used to identify conserved regions that play functional or structural roles in collections of biosequences that are assumed to be related. The most common way to align multiple sequences is progressive alignment. In this approach, the most similar pair of sequences is aligned first, and then more distant sequences are added progressively .[23]. The method has the important characteristics that it can detect discontinuous patterns which equips it with the potential to successfully handle more complex structures with non-concatenative properties.

2 Theoretical Considerations

In this section we briefly survey the challenges that unsupervised morphological segmentation methods face when coping (a) with the subject of the task, the language; (b) with the methodological issues and implementations; and (c) with the evaluation of the segmentation results. The issues discussed in this section are then instantiated in the empirical part in Section 3.

2.1 Language challenges

The utmost aim of unsupervised, knowledge-free algorithms is to analyse any given language. However, world languages display a variety of morphological processes and phenomena, so that designing one system that can successfully handle them is a very ambitious challenge. Depending on the predominance of particular morphological process, a language can be classified as agglutinative, inflectional (with the subclass of introflectional), isolative, or polysynthetic. In general, languages are of mixed morphological types, so that only a method that can principally handle any language type can also handle any morphological process that might occur in a language.

Each morphological process poses a different challenge to unsupervised segmentation. Whereas in polysynthetic languages the task of morphological segmentation overlaps to a great deal with word segmentation, strictly isolative languages like Chinese are not relevant for morphological segmentation, since they do not contain any morphemes that could be split. Especially agglutination and inflection thus challenge the unsupervised language analysis, each of them having properties that either complicate, or alleviate the morphological segmentation task.

Agglutinative languages exhibit clear boundaries between morphemes which may be advantageous for unsupervised segmentation. On the other hand, several affixes are often agglutinated to one stem, which lowers the frequency of occurrence of the same word forms in the corpus, and thus has negative implications for finding contextually similar words. As stated earlier, present unsupervised methods perform best on one slot linear morphology.

Inflectional languages have typically smaller number of affixes which can be added to a root, due to accumulation of several functions on one affix and frequent syncretism among inflected forms. However, stem alternants and affix allomorphy are also more frequent phenomenon in these languages, which makes it more difficult to determine where the segmentation point should be and whether formally similar word forms are also close morphologically. A special case are introflective languages with so called non-concatenative morphology like e.g. Hebrew or Arabic, which often contain discontinouos morphemes both as stems (e.g. root consonants) and inflections (introflection).

As mentioned above, phenomena that characterize one language type are typically present also in languages assigned to another predominate type. For example, German, the languages on which we demonstrate our method and the challenges to the unsupervised segmentation, is usually classified as a an inflectional language. In addition to the inflection ,however, German features other morphological processes. Similarly to English, grammatical relations between words can be expressed both through inflectional suffixes, e.g., *Werthers Leiden* 'Werther's suffering', or through prepositions, which is

typical for isolative languages, *das Leiden von Werther* 'the suffering of Werther'. There are also agglutinative phenomena like in the word *Kind-er-n* child-PL-DAT, where each affix carries only one grammatical function; -er: plural -n: dative. Moreover, German compounding that can result in very complex words is a property that links German with polysynthetic languages. German irregular verbs (e.g. *singen* - *sangen* , sing.PRS-PL-sing.PST-PL) represent an example of introflection in this language, since grammatical category (here tense) changes depending on the vowel inserted into the discontinuous stem morpheme. Circumfixation, another non-concatenative phenomenon, is present in German as well, e.g. in *ge-lauf-en* PTCP-run-PTCP, where the circumfix ge-en marks the participle form.

Methods that aim at an adequate and complete analysis of any given language should handle successfully all morphological structures. However, as stated also by [15, p. 310, p.332], unsupervised methods tend to exhibit an implicit or explicit bias towards a certain kind of languages. Many unsupervised algorithms for morphological analysis assume concatenative morphology, and design their methods and data structures in ways that are suited to describe such phenomena (e.g. [25, 26, 10, 4]).

In the empirical part of this paper we present a method that tries to avoid such bias by applying MSA that had been shown to be able to deal with both concatenative and non-concatenative morphology. The method differs from the previous approaches in that it thrives to be language independent (cf. [24] with strong language specific bias for Arabic, or [8] for stem variation in German) and that it focuses on morphological segmentation (cf. [3] who addresses morphology induction with very low F-scores (below 0.10) for the non-concatenative Arabic). Our approach is similar to features-and-classes method of [7] in its attempt to design a general system that can address the whole range of morphological phenomena in any given language.

In addition to the typological division of languages and morphological processes outlined above, another relevant distinction is between inflectional and derivational morphology. Inflection modifies a word to create new word forms that express different grammatical categories (e.g. number, case, tense, etc.). Regular inflection is typically very productive and therefore easier to detect automatically. Derivation, on the other hand, is a less productive process that creates new lexemes out of existing bases. It involves change in the core meaning and often also change in the word class. Lower productivity and the involved change of meaning are two aspects of derivational morphemes that make them more difficult to detect in an unsupervised manner. Some methods therefore resign on this aim completely, for example [25, 26]. [13, 14], on the other hand, decompose word forms into stems and suffixes, using the Minimum Description Length (MDL) principle, and groups them into signatures, each is a structure that denote a set of stems that can co-occur with a set of affixes. This method handles inflectional and derivational morphology without making a clear distinction between them. However, it is, again, restricted to concatenative morphology. The qualitative analysis of the data produced by the system described below shows that our approach can segment more complex inflectional and derivational morphemes (see also Table 3).

2.2 Algorithmical and resources challenges

Morpheme is the smallest unit of language that carries meaning, i.e. it is a particular form bound to a particular meaning or function. The mapping between them is not always one-to-one (cf. allomorphy) and the form does not need to be linear (see above). However, given this twofold nature of a morpheme it is obvious that unsupervised methods ignoring the meaning/function aspect of morphemes are doomed to fail: Languages abound of strings that formally overlap but do not have the status of a morpheme.

A particular challenge related to unsupervised, knowledge free morphological analysis thus is how to approximate the meaning. One possibility that is exploited also by the method presented in this paper is to use context. In line with the Distributional Hypothesis [16] words that appear in the same context are semantically similar. However, in order to compute distributional similarity reliably, a sufficient amount of contexts in which the word forms occur is needed. Consequently, most current unsupervised methods are very resource intensive, in that they require corpora of a very large size. The need for huge corpora is so acute that even a corpus of 500000 running forms is considered small by some methods [9]. For resource rich languages, large corpora and possibly also training sets for supervised algorithms are not a problematic issue. However, a field that could profit substantially from unsupervised language analysis are resource poor and/or endangered languages for which sometimes only small, unannotated corpora are available. Clearly, such settings do not provide enough input for context based methods to reliably detect distributionally and thus semantically similar words.

In parallel to the typological problem described in the previous section, where a failure of a method to deal with a particular language type also means a failure to deal with some morphological features in a given language (since language types are mixed), the data sparseness problem described above does not affect the performance of the algorithms only on small corpora, but also on corpus low frequent word forms in a large corpus. The system presented in this paper attempts to find an solution that delivers adequate analysis also for corpus low frequent words.

2.3 Evaluation challenges

The most widely used evaluation method is the automatic comparison of the computed results against adequate linguistic reference, i.e. the gold standard. The alternative, a direct manual evaluation by the language expert(s) is both time and work consuming and unrealistic in many settings. Depending on the task (and also the available gold standard), various evaluation methods have been proposed. Their overview can be found in [28]. In the area of morphological segmentation, the most straightforward evaluation is the calculation of how well the automatically detected segmentation boundaries correspond to the morpheme boundaries in the gold standard (e.g., [6, 20]).

The first challenge to the automatic segmentation evaluation is the availability of adequate reference analyses, the gold standard, which typically exist only for resource rich languages. The MorphoChallenge competition (since 2005) provided gold-standard evaluation data for English, German, Finnish, Arabic, and Turkish and for task-based Information Retrieval evaluation data for English, German, and Finnish. Though the MorphoChallenge series without doubt greatly supported the research in unsupervised

morphological analysis and contributed to the evaluation standardization and comparability, it is also evident that given the diversity of world languages, the offered sample cannot be viewed as representative (see also [15, p.335]). It has been acknowledged (see e.g. [28]) that unsupervised methods cannot come with results that exactly correspond to those designed by linguists. However, it is not only the limitations of the unsupervised methods but also of the gold standards quality that challenges the evaluation. Their reference analyses often do not correspond to complete linguistic analyses. One typical problem are derivational affixes. Whereas most gold standards for morphological segmentation contain all or close to all boundaries separating inflectional affixes, segmentation of derivational affixes can be missing or incomplete. Consequently, a method that is able to detect boundaries also between roots and derivational affixes can be disadvantaged compared to a method focusing solely on inflection when compared to such a gold standard. Even more intriguing is the problem of introflection. Changes on stems are typically not grasped by the gold standards. In German, word forms like singen and sangen are analysed only with respect to their inflectional suffix, i.e. *sing-en* and *sang-en*. A method that correctly identifies the vowel change within a discontinouos stem and performs the analysis as *s-i-ng-en* and *s-a-ng-en* is penalized because the additional splits are scored as incorrect (cf. Table 3).

In addition to the challenges related directly to the gold standards, there are challenges related to how the evaluation is performed by individual authors. Low frequent word forms (which can mean up to ten occurrences in this context, cf. [26], or all other words except the most frequent ones, cf. [10, 8], are often excluded from either the analysis itself, or from the evaluation, or from both. Moreover, the unsupervised algorithms often perform poorly on the most frequent words. As an example, the algorithm presented in [2] delivers worse results without the trimming of the word forms with a corpus frequency above 0.01% of the total token count. The authors argue that these tend to be function words that are of little interest for morphological analysis. The evaluation of the algorithms thus often differs not only with respect to which gold standard is used and what its properties are, but also with respect to the corpus portion that had been analysed and reported.

3 Empirical Instantiations

When designing the unsupervised segmentation system described below, we carefully considered the challenges outlined in the theoretical part. Given the two-sided nature of a morpheme, we decide for a system that takes into account not only morpheme's formal aspects, but also its meaning/function. In order to approximate meaning, we decided to exploit the distributional hypothesis following the thesis that words that occur in similar contexts have also similar meaning/function. Since such approach can be successfully applied only on word forms with sufficient corpus frequency, we designed a system comprising of two methods: One is using context similarity directly, the other indirectly through utilizing the results of the first method.

As already mentioned, various morphological processes are involved in word form construction. In order to capture this morphological variance, we based the system on MSA that has the potential to address any of the morphological processes. There are

only few biologically inspired methods reported for the task unsupervised of morphological segmentation: [11, 18] employ genetic algorithm to obtain the optimal solution within the space of all possible word segmentation into stems and suffixes. These methods use fitness functions which can be viewed as simplified forms of MDL: They seek the absolute minimum of characters [18] or elements [11] in the sets of stems and suffixes, that describe the language, rather than using the information-theoretic criterion, which is based on conditional probabilities, as in [13]. [12] enhance the above idea to detect derivational paradigms, with a strategy that takes into account a property of the language, i.e., the fact that different stems may be combined with the same set of suffixes. The method first generate hypothesized stems and suffixes from a list of words; for each stem all possible paradigms are detected, i.e., the sets of suffixes repeated for that stem. Binary chromosomes represent solutions, where each gene (index in the chromosome) encodes a hypothesized stem or suffix. The genetic algorithm is then applied to an initial population of randomly generated chromosomes. As can be observed also here, these methods focus their efforts on analyzing concatenative morphology, identifying suffixation patterns. Previous work utilized MSA for morphological segmentation, but handled this task differently. [27] aligns orthographically similar words, and uses third-party analysis [21] as a guide, to search for a set of segmentation columns. It determines its segmentation decisions by maximizing the F-score against the analysis of the third-party system. A more closely related work is presented in [19], where semantically related words are used to identify patterns which are assumed to be morphemes. However, similarly to other approaches relying on contextual information, also this approach analyses only those word forms in the corpus that appear with sufficient frequency.

4 The First Method M_1

M_1 is based on the idea that morphologically related words are both formally and semantically similar. We assume that recurring patterns within such words correspond to the morphological relations among them. We identify overlapping patterns within such word with the assistance of MSA, and insert predicted morpheme boundaries in those words accordingly.

In the first step, distributionally and othographically similar word forms are extracted and clustered into sets of presumably morphologically related word forms according to a method described in [19]. In the second step, patterns are extracted from the sets using MSA. Word forms in these sets are aligned using progressive alignment: First, the two most similar sequences are aligned and then less similar ones are added in a cumulative way to construct the final alignment. In the context of morphological segmentation, selected sets of distributionally and orthographically similar words are treated as sequences that are to be aligned. The first sequence of the alignment is the input word, and the similarity criterion means, in this case, similarity of a related word to the input word. The alignment method is based on the one appears in the BioJava package [17], modified for our purpose. Table 1 demonstrates the alignment set for the word *umgedreht*. The "−" signs indicate gaps which are inserted during the alignment process to unify the lengths of the sequences.

Table 1. An example for an alignment of the word form *umgedreht* and its related word forms.

umgedreht---
abgedreht---
um--dreht---
um--drehte--
...
umzudrehe--n
umgesied-elt
um--drehe--n
...

Next, M_1 compares the aligned sequences to find a pattern which matches the alignment best: Identical fragments are extracted from pairs of aligned sequences, and are considered as candidate patterns for this alignment. Each candidate pattern is stored with the number of corresponding sequences with which it matches. In the example above the pair of aligned sequences constructed from the word forms *umgedreht* and *abgedreht* generates the candidate pattern `-gedreht`, whereas the pair of aligned sequences consisting of the word forms *umgedreht* and *umdrehen* generates the candidate pattern `um-dreh` that contributes to the correct and complete analysis of the word form `um-ge-dreh-t`. Each candidate pattern $pattern_i$, of the alignment set is given a score which balances between the relative frequency of the pattern in the given alignment and the length of the pattern, and is calculated as follows:

$$score(pattern_i) = \frac{2}{\frac{count(pattern_i)}{\sum_j count(pattern_j)} \log(size) + \frac{1}{length(pattern_i)}} \; . \qquad (1)$$

Here, $count(pattern_i)$ is the number of aligned sequences which match this candidate pattern, $size$ is the number of sequences participating in this alignment and $length(pattern_i)$ is the number of characters which $pattern_i$ consists of. Patterns are ranked based on their scores, and the pattern that got the highest score is selected as the one that describes best the members of the alignment set, and the word forms which formed that alignment set are segmented accordingly. This candidate segmentation for each word from is recorded along with the respective score. A word form may be a member of several alignment sets since it can match the condition of both distributional and form similarity for more than one input word form, and it can be an input word form itself. Therefore a word form can have several candidate segmentations from the different alignments, some of which can be identical. To select the best segmentation for each word form, the scores recorded for each candidate segmentation are tallied, and a ranked list of possible segmentations for each word form is constructed based on those scores.

The method was applied on a corpus of three million German sentences obtained from the Wortschatz collection[1] at the University of Leipzig (Germany).[2] Overall, out

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

[2] These sentences were used in MorphoChallenge comptetions.

of 1294071 M_1 was able to analyse 196852 (15.2%) word forms, of which 58213 were found in CELEX [1] which is used as a gold standard.

Since M_1 returns a ranked list of segmentation options for each word, we report the top-1, -2 and -5 results. The results are summarized in Table 2 and show the precision (P), recall (R), and F-measure (F) values for each of these cases.

Table 2. Results for M_1

Top-n	P	R	F
1	0.48	0.46	0.47
2	0.57	0.56	0.56
5	0.60	0.59	0.59
Baseline	0.22	0.49	0.30
Morf.	0.60	0.43	0.50

The results were compared to a baseline which assigns segmentation points to the input word forms randomly. Our method performs well above the baseline which achieved F-score of 0.30. The results were also compared to Morfessor [5] which represents the current state of the art. The comparison shows that the presented method achieves good results that for top-1 are close to the state of the art with a potential for further improvement as indicated by the top-2 and top-5 results. It should be pointed out that the top-2 and top-5 analyses hypotheses do not present mutually exclusive solutions. Instead, they typically comprise several solutions that differ in how close they are to the complete linguistic analysis.

A qualitative analysis of the data confirms that the method can deal with different morphological processes that are sometimes not grasped by Morfessor, or by the gold standard, or by both. Table 3 gives overview of such examples.

Table 3. Examples of morphological processes analysed by M_1 with their corresponding F-scores when compared to the gold standard: D - derivation, I - inflection, /Intr/ - introflection, Cfix - circumfixation, Aggl - agglutination, P/C - polysynthesis/compunding.

complete analysis	Process	M_1	Morfessor	gold standard
alban-isch	D	alban-isch (1.0)	albanisch (0.0)	alban-isch
hebrä-isch	D	hebrä-isch (0.0)	hebräisch (1.0)	hebräisch
Raps-öl	P/C	Raps-öl(1.0)	Rapsöl (0.0)	Raps-öl
ge-wähl-t	Cfix	ge-wähl-t (1.0)	gewählt (0.0)	ge-wähl-t
k/a/nn	/Intr/	k-a-nn (0.0)	kann (1.0)	kann
zu-ge-ruf-en	D+Cfix	zu-ge-ruf-en (1.0)	zu-gerufen (0.3)	zu-ge-ruf-en
Stief-kind-er-n	D+Aggl	Stiefkind-er-n (0.5)	Stiefkind-er-n (0.5)	Stief-kind-ern

5 The Second Method M_2

M_1 analyses word forms for which both distributionally and formally similar word forms could be retrieved. This approach requires enough characteristic contexts to create a reliable contextual representation of a word from. Words forms with low frequency can therefore achieve only inaccurate representations and the degree of semantic relatedness among them is typically rather weak. Consequently, the probability that a morphologically related words would be among them is also lower.

Method M_2 presented in this section was designed to handle the word forms with such context constraints. In order to avoid an approach that would take into account solely the form aspects of morphemes, the meaning/function based segmentation results from the first method were utilized to compute the segmentation of the so far unanalysed words.

For each unanalysed input word (focus word), M_2 first collects a list $\{w_k\}$ of previously analysed words that are formally similar to it. Form similarity between the focus word and w_k is calculated as $1 - d_k$, where d_k is Needlman-Wunsch distance [22] with affine gap penalties, normalized to the range [0,1]. $\{seg_{kl}\}_{l=1}^{5}$ is formed by retrieving the top-5 analyses for each w_k. The focus word form is then compared to each seg_{kl} to find matching segments and M_2 generates segmentation hypotheses h_{kl} for the focus word.

In our experiments with this method we considered several parameters. The implementation with the so far best results included (a) the degree of form similarity between the focus word and each of the words in $\{w_k\}$, as described above; (b) the coverage of segments found in the focus word with respect of to segments in seg_{kl}, that is, the ratio of the segments found in the focus word and the segments of seg_{kl}. Let $\{s_m\}$ be the set of segments of a given segmentation hypothesis seg_{kl} of w_k, and let $\{s_n\} \subset \{s_m\}$ be the set of segments discovered in the focus word, then the ratio between the sizes of these two sets is used as a measure of segments coverage. The score for a single segmentation hypothesis of a focus word then is:

$$score(h_{kl}) = \frac{|\{s_n\}|}{|\{s_m\}|} \times (1 - d_k) \; . \tag{2}$$

The results of this experiment are presented in Table 4.

Table 4. Results for M_2

Top-n	P	R	F
1	0.49	0.44	0.46
2	0.62	0.56	0.59
5	0.68	0.63	0.65
Baseline	0.21	0.50	0.29
Morf.	0.74	0.52	0.61

In our future work, we want to include other parameters in our experiments and investigate the potentials of various parameter combinations to optimize both the top-1 and top-n results. One possibility would be including a "segment validity parameter" that can be computed as a function of segment frequency among different seg_{kl}s.

6 Overall Results and Evaluation

The evaluation of the whole corpus including words analyzed by both M_1 and M_2 is presented in Table 5.

Table 5. Results for the whole corpus

Top-n	P	R	F
1	0.48	0.46	0.47
2	0.59	0.56	0.57
5	0.64	0.61	0.63
Morf.	0.67	0.48	0.56

The results show that the system delivers useful results when applied on data with different degree of sparseness. Though top-1 analyses are still subject to improvement, the top-5 results show that the method can achieve promising results. The qualitative analysis of the results confirms (see also Table 3) that the evaluation is negatively affected by some properties inherent to the gold standard. Introflective aspects of German that include e.g. stem vowel changes in the conjugation of irregular or auxiliary verbs or in pluralization of nouns are not captured by the gold standard but are often analysed by our system. Consequently, segmentation boundaries that are actually correct are scored as false positives due to their absence in the gold standard. Similarly, not all derivation morphemes are segmented in the gold standard either (see the examples Hebräisch and Albanisch in Table 3). Due to these deficits in the gold standard, the qualities of the presented system that distinguish it from other approaches sometimes fall short compared to methods that deliver results more conform to the (imperfect) gold standard. It can be however assumed that its actual performance is higher than the evaluation reveals.

It would be probably unrealistic to expect that any existing gold standard of sufficient size for the evaluation of unsupervised methods could contain a complete and perfect annotation. The more important it then seems for the comparability of the results that the specifics of the gold standard used for evaluation would be at least partly as well described as the evaluation method itself. It would be further useful if the quantitative analysis was accompanied by at least a brief qualitative analysis of the method's output, so that the reader can get insights into its scope. As an example, compared to methods that are biased or designed to perform (only) the segmentation of inflectional affixes, systems such as the one presented in this paper may not achieve equally high results on the regular and frequent inflectional phenomena, but might be able to address a larger scope of morphologically different processes.

7 Conclusions

In the first part, the paper surveyed the theoretical context and challenges of unsupervised, knowledge free morphological segmentation. In the second part it described a system grounded in the theoretical considerations and presented its result on German. The method utilizes MSA to analyse formally and distributionally similar words, and uses these context based results to assist the analysis of less frequent words. The results show that the method can handle a broad range of morphological processes in a quality close to the present state of the art approaches and has potential for further improvement.

References

1. Baayen, R.H., Piepenbrock, R., Gulikers, L.: The CELEX lexical database (release 2). CD-ROM (1995)
2. Baroni, M., Matiasek, J., Trost, H.: Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In: Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning. pp. 48–57. Association for Computational Linguistics (July 2002)
3. Bernhard, D.: Morphonet: Exploring the use of community structure for unsupervised morpheme analysis. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Penas, A., Roda, G. (eds.) Multilingual Information Access Evaluation I. Text Retrieval Experiments, LNCS, vol. 6241, pp. 598–608. Springer (2010)
4. Bordag, S.: Unsupervised and knowledge-free morpheme segmentation and analysis. In: Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. LNCS, vol. 5152, pp. 881–891. Springer (2008)
5. Creutz, M., Lagus, K.: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Tech. Rep. Report A81, Helsinki University of Technology (March 2005)
6. Creutz, M., Lindén, K.: Morpheme segmentation gold standards for finnish and english. Publications in Computer and Information Science, Report A 77 (2004)
7. De Pauw, G., Wagacha, P.W.: Bootstrapping morphological analysis of gıkuyu using unsupervised maximum entropy learning. In: Proceedings of the Eighth Annual Conference of the International Speech Communication Association. (2007)
8. Demberg, V.: A language-independent unsupervised model for morphological segmentation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 920–927. Association for Computational Linguistics (June 2007)
9. Fisher, D., Riloff, E.: Applying statistical methods to small corpora: Benefitting from a limited domain. In: Probabilistic Approaches to Natural Language, a AAAI Fall Symposium. pp. 47–53 (1992), technical Report FS-92-04
10. Freitag, D.: Morphology induction from term clusters. In: Proceedings of the Ninth Conference on Computational Natural Language Learning. pp. 128–135. CONLL '05, Association for Computational Linguistics (2005)
11. Gelbukh, A.F., Alexandrov, M., Han, S.: Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model. In: Sanfeliu, A., Trinidad, J.F.M., Carrasco-Ochoa, J.A. (eds.) Progress in Pattern Recognition, Image Analysis and Applications, 9th Iberoamerican Congress on Pattern Recognition, CIARP '04. LNCS, vol. 3287, pp. 432–438. Springer (2004)

12. Gelbukh, A.F., Sidorov, G., Lara-Reyes, D., Chanona-Hernández, L.: Division of spanish words into morphemes with a genetic algorithm. In: Natural Language and Information Systems, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008. LNCS, vol. 5039, pp. 19–26. Springer (2008)
13. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics 27(2), 153–198 (2001)
14. Goldsmith, J.: An algorithm for the unsupervised learning of morphology. Natural Language Engineering 12(04), 353–371 (2006)
15. Hammarström, H., Borin, L.: Unsupervised Learning of Morphology. Computational Linguistics 37(2), 309–350 (2011)
16. Harris, Z.S.: Distributional Structure. In: Fodor, Jerry A. and Katz, Jerrold J. (ed.) The Structure of Language: Readings in the Philosophy of Language, pp. 33–46. Prentice-Hall (1964)
17. Holland, R.C.G., Down, T.A., Pocock, M.R., Prlic, A., Huen, D., James, K., Foisy, S., Dräger, A., Yates, A., Heuer, M., Schreiber, M.J.: BioJava: an open-source framework for bioinformatics. Bioinformatics 24(18), 2096–2097 (2008)
18. Kazakov, D.: Unsupervised Learning of Naïve Morphology with Genetic Algorithms. In: Daelemans, W., van den Bosch, A., Weijters, A. (eds.) Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks. pp. 105–112 (1997)
19. Kirschenbaum, A.: Unsupervised segmentation for different types of morphological processes using multiple sequence alignment. In: Dediu, A.H., Martín-Vide, C., Mitkov, R., Truthe, B. (eds.) Proceedings of Statistical Language and Speech Processing, SLSP. LNCS, vol. 7978, pp. 152–163. Springer (2013)
20. Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E., Saraclar, M.: Unsupervised segmentation of words into morphemes-challenge 2005: An introduction and evaluation report. In: Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes (2006)
21. Monson, C., Hollingshead, K., Roark, B.: Probabilistic ParaMor. In: Working Notes for the CLEF 2009 Workshop (2009)
22. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48(3), 443–453 (1970)
23. Notredame, C.: Recent progresses in multiple sequence alignment: a survey. Pharmacogenomics 3(1) (2002)
24. Rodrigues, P., Ćavar, D.: Learning arabic morphology using statistical constraint-satisfaction models. In: Benmamoun, E. (ed.) Perspectives on Arabic Linguistics XIX, pp. 63–75. John Benjamins (2007)
25. Schone, P., Jurafsky, D.: Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. In: Proceedings of the 4th Conference on Computational Natural Language Learning-Volume 7. pp. 67–72. Association for Computational Linguistics (2000)
26. Schone, P., Jurafsky, D.: Knowledge-free induction of inflectional morphologies. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. NAACL '01, Association for Computational Linguistics (2001)
27. Tchoukalov, T., Monson, C., Roark, B.: Morphological Analysis by Multiple Sequence Alignment. In: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009. LNCS, vol. 6241, pp. 666–673. Springer (2010)
28. Virpioja, S., Turunen, V.T., Spiegler, S., Kohonen, O., Kurimo, M.: Empirical comparison of evaluation methods for unsupervised learning of morphology. Traitement Automatique des Langues 52(2), 45–90 (2011)