4

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 25–31
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4513

Unsupervised morphological analysis of small corpora: First experiments with Kilivila

Amit Kirschenbaum[a], Peter Wittenburg[b], and Gerhard Heyer[a]

[a]*University of Leipzig*
[b]*Max Planck Institute for Psycholinguistics, Nijmegen*

Language documentation involves linguistic analysis of the collected material, which is typically done manually. Automatic methods for language processing usually require large corpora. The method presented in this paper uses techniques from bioinformatics and contextual information to morphologically analyze raw text corpora. This paper presents initial results of the method when applied on a small Kilivila corpus.

1. INTRODUCTION. Unsupervised approaches to language processing attempt to discover the structure of language based on machine-learning techniques applied to unannotated text. Present unsupervised methods are typically developed for large corpora and are not suitable for small corpora of a few hundred thousand or even just thousands of words. However, this is the typical size of corpora in language documentation (see Klamer, this volume; Holton, this volume), where the potential of unsupervised learning methods has already been acknowledged:

> "Basic linguistic descriptions of lexicon and grammar made on the basis of transcribed recordings still form an important component of language documentation, however, and with the realization that languages are disappearing at a far faster rate than linguists can document them, it is natural to look for ways of making this process less labor-intensive." (Hammarström & Borin 2011)

In particular, unsupervised methods for morphological analysis aim to learn the internal structure of words from a raw text corpus of a given language, and this typically means segmenting words into morphemes.

Unsupervised morphological analysis can be traced back to an algorithm introduced by Harris (1955, 1967), and later improved by Hafer & Weiss (1974). The algorithm detects morpheme boundaries as a function of the number of distinct letters that follow, or precede, a letter sequence which is part of a word (Letter Successor Variety). If a peak is reached in that number, then it would probably be due to a morpheme boundary.

A different approach incorporates the Minimum Description Length (MDL) principle (Rissanen 1978), which is based on information-theoretic grounds. This principle follows

the idea that regularities in data can be used "to describe it using fewer symbols than the number of symbols needed to describe the data literally" (Grünwald 2007). It therefore seeks to minimize a cost function which is the sum of the description length[1] of the model explaining the data, and of the description length of the data, when encoded with this model. Goldsmith (2001, 2006) used MDL to construct lists of stems, suffixes, and signatures, i.e. structures that indicate which stems may appear with which suffixes. Another MDL based method is presented by Creutz & Lagus (2002). In this method, the model is a lexicon of morphs, which may either be prefixes, suffixes, or stems. The data is encoded by sequences of pointers to the lexicon. Each input word is segmented in various ways. For each way, the cost is evaluated, and the segmentation which achieved the minimum cost is selected. This method is recursive, and each new morph may be subject to further splitting.

Another type of algorithms also uses contextual features as part of the morphological segmentation process. Freitag (2005) utilizes local co-occurrence information to create clusters that correspond roughly to syntactic classes. The method then induces affix transformation rules which express relations between clusters and show possible affixation patterns. Further description of other methods devoted to unsupervised morphology can be found in a detailed survey by Hammarström & Borin (2011).

The purpose of the method presented in this paper is to help alleviate the situation in language documentation as described above by providing a way to automatically segment small corpora on the morphological level in order to facilitate lexical and morphological analysis and description.

The present method is (in principle) language independent and should work for languages with various properties (e.g. with both concatenative and non-concatenative morphology, various word orders, etc.). It employs word-distributional similarity and sequence alignment, a technique borrowed from bioinformatics in which protein or DNA sequences are compared, and similar regions among them are then identified.

For the purpose of this paper, the method was applied to a small corpus of Kilivila (Senft 1983–1997), an Austronesian language spoken by the Trobriand Islanders of Papua New Guinea.[2] The examples in the following section come from this corpus as well.

2. METHOD. The method starts with computing a word co-occurrence model to discover distributional similarities between words. The model is represented in a high-dimensional vector space, where every word in the corpus is associated with a *context co-occurrence vector*. The context vector for a word *w* is defined by words which co-occur with *w* within a sentence context.[3] To reduce noise, the context vector for *w* is represented only by significant co-occurrences of *w*. To extract the significant word co-occurrences we utilize the method described in Quasthoff & Wolff (2002). This method is based on comparing the expected number of joint occurrences of two words in a corpus, under the independence assumption, to their actual number of co-occurrences in that corpus. Examples of some

[1] The length is measured in bits.

[2] We would like to thank Prof. Gunter Senft for providing the corpus and the accompanying linguistic analysis.

[3] The term *co-occurrence*, henceforth, will refer to joint occurrence of two words within sentential context.

co-occurrences (and their morphological analyses) of an input word in Kilivila *bukuninamsisi* [2FUT-think-PL][4] from our test corpus are given in Table 1.

In the next step, similarity relations between words are computed by comparing their context vectors. The underlying rationale is based on the distributional hypothesis by Harris (1968) according to which words with similar distributional properties (i.e., contexts) tend to be semantically similar. We employ the method described in Bordag (2008) to compare context vectors and obtain distributionally similar words. Table 2 shows some of the distributionally similar words for the word *bukuninamsisi*.

mankawa	[DEM-DEM-CP.thing]
ekau	[3PRS-take]
yegulaga	[I-Emph-Emph]
sogu	[friend-my]
isiligaga	[3PRS-important]
makaukweda	[our-veranda]
bibwadi	[3FUT-be.possible]
beya	[here,there,this]
bagisi	[1FUT-see]
tabu	[taboo]
...	...

TABLE 1: Co-occurrences examples for the word *bukuninamsisi* [2FUT-think-PL]

lalilivali	[1PST-tell]
bukusisusi	[2FUT-be-PL]
lumkola	[feeling]
bukukanukwenusi	[2FUT-lie.down-PL]
biboda	[3FUT-be.good]
ibubulisi	[3PRS-work-PL]
nanomi	[mind-your]
biyapu	[3FUT-be.good.and.bad]
bukulilolasi	[2FUT-walk-PL]
evagisi	[3PRS-make-PL]
...	...

TABLE 2: Examples for distributionally similar words of the word *bukuninamsisi* [2FUT-think-PL]

The set of distributionally similar words of w is then filtered based on edit distance (Needleman & Wunsch 1970) from w. The resulting target set consists of words which are both distributionally and orthographically similar to w. Orthographically close words

[4] Abbreviations: 1, 2, 3 – 1st, 2nd, 3rd person; CP – Classificatory Particle; DEM – demonstrative; Emph – emphatic; FUT – future; PL – plural; PRS – present; PST – past

may be derivations or inflections of one stem, or words from one word class sharing a set of morpho-syntactic features (e.g., verbs in the same tense).

The words from the target set are aligned using a *multiple sequence alignment* (MSA) method. Multiple sequence alignment aims to discover functional, structural, or evolutionary relationships among a set of biosequences by searching for character patterns in these biosequences. The alignment process inserts "gap" into the sequences allowing equivalent characters from different sequences to be positioned in the same column.

We employed BioJava (Holland et al. 2008), a bioinformatics toolkit, to perform the alignment. The strategy used was, first, to align *w* and its orthographically most similar word from the target set and then to gradually align less similar words in a cumulative fashion. The result is a set of aligned words from which one or more segmentation patterns can be extracted based on character overlap in the aligned sequences. Table 3 demonstrates a part of the alignment of the word *bukuninamsisi* and its target set. '–' signs mark the "gaps" inserted in words during the alignment process.

b	*u*	*k*	*u*	–	–	–	*n*	*i*	*n*	*a*	–	*m*	*s*	*i*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*n*	–	–	*a*	–	*m*	*s*	*i*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*n*	*o*	*k*	*a*	*p*	*i*	*s*	*i*	*s*	*i*
–	–	*k*	*u*	–	–	–	*n*	*a*	*n*	*a*	–	*m*	*s*	*a*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*l*	*i*	*g*	*e*	–	*m*	*w*	*e*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*n*	*u*	*k*	*w*	*a*	*l*	–	*i*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*s*	*i*	–	–	–	–	*s*	*u*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*m*	*a*	–	–	–	–	*s*	*i*	*s*	*i*
b	*i*	*t*	*a*	–	–	–	*n*	*i*	*n*	*a*	*m*	–	–	–	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*t*	*e*	*m*	*a*	*l*	–	–	*i*	*s*	*i*
...																

TABLE 3: Part of the alignment for the word *bukuninamsisi* and its target set

The patterns are scored according to their length in characters and according to the number of words in the target set which they match. We record the pattern with the highest score as a possible segmentation of the words in the target set. The pattern extracted for the set presented in Table 3 is *buku-si*, which encodes [2FUT-VERB-PL], and words that match this pattern would be segmented according to it. There is no restriction for the form of that extracted pattern, and it is indeed possible that patterns would also include e.g., infixes.

However, the word *w* can also be a member of target sets of different words {*w'*} for which different patterns might be recorded as possible segmentations. Consequently, for a given *w*, several segmentation patterns may be recorded, and each of them may appear more than once. Hence, we get a weighted set of possible patterns for morphological segmentation of *w*, and we select (in the current implementation) the pattern with the highest frequency.

3. EVALUATION. We applied the method on a small corpus of narrations in Kilivila (Senft 1983–1997). The corpus consists of ca. 13,000 words. The corpus contains

morphological annotations constructed by an expert and thus serves as our "gold standard" reference.

The evaluation method compares the segmentation decisions derived from our method for each word to the actual segmentations in the reference. Segmentation points that were marked by the method and correspond to actual morphological boundaries are called true positives (*tp*). Segmentation points which were identified by the method but do not correspond to morphological boundaries are called false positives (*fp*). Morphological boundaries which were not detected by the method are called false negatives (*fn*). Precision and recall are then calculated for each word, based on the amount of segmentation points in each of the above categories.

$$Precision_w = \frac{\#tp}{\#tp + \#fp} \qquad Recall_w = \frac{\#tp}{\#tp + \#fn}$$

Thus, the precision measures the portion of segmentation points which are correct, out of the reported segmentation points, and recall measures the portion of those correctly found segmentation points out of the actual segmentation points. Both measures reach a maximum of 1 when there are no mistakes in segmenting the word by the method. The value decreases when there are prediction errors, i.e. redundant segmentation points in the case of precision, or missed ones in the case of recall. When no segmentation point is identified correctly, the value of these measures is 0. The average precision (P) and recall (R) are then calculated based on the results for each word.

Table 4 summarizes the results. The first line is our baseline, which randomly assigns segmentation points as morpheme boundaries. The second line presents the results for applying the method on the whole corpus. The third line presents the evaluation results for using the method after setting a reliability threshold on the derived patterns.

METHOD	P	R
Random	0.22	0.44
Unsupervised	0.381	0.569
Unsupervised+thresh	0.682	0.133

TABLE 4: Evaluation results

4. FUTURE WORK. The method presented here still requires much exploration. We plan to experiment further with the ways of extracting patterns from target sets and determining the final segmentation, and we also plan to experiment with different corpus sizes. The present version of the method assumes existing sentence boundaries; however, we plan to experiment with sentence independent context windows as well.

A future version of this method should also be able to derive the morphology of the analyzed language, in the sense of supplying the user with generalizations regarding, for example, inflectional and derivational paradigms. This method is intended to be a component in

a larger framework of automatic annotation which would consist of both unsupervised and supervised algorithms.

The unsupervised module of the system would attempt to compensate the data sparseness problem by using linguistic information of various sources (morphological, parts of speech, semantic levels) and by taking advantage of the interaction between these levels. As a result, the module would produce suggestions for linguistic analyses on the three levels, which the annotator can manually correct. The supervised module would then train a model on the corrected data and would produce the final annotation of the corpus.

This system of interactive annotation is planned to be integrated into existing and widely used environments such as ELAN[5] (Wittenburg et al. 2006) or LEXUS[6] (Kemps-Snijders et al. 2006) in order to make the annotation process more efficient.

REFERENCES

Bordag, Stefan. 2008. A comparison of co-occurrence and similarity measures as simulations of context. In *Proceedings of the 9th International Conference Computational Linguistics and Intelligent Text Processing (CICLing)*, 52–63. Springer.

Creutz, Mathias & Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, 21–30. Association for Computational Linguistics.

Freitag, Dayne. 2005. Morphology induction from term clusters. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 128–135. Association for Computational Linguistics.

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

Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(4). 353–371.

Grünwald, Peter D. 2007. *The Minimum Description Length Principle*. Cambridge, Mass: MIT Press.

Hafer, Margaret A & Stephen F Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10. 371–385.

Hammarström, Harald & Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2). 309–350.

Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31(2). 190–222.

Harris, Zellig S. 1967. Morpheme boundaries within words: Report on a computer test. *Transformations and Discourse Analysis Papers* 73. Philadelphia: University of Pennsylvania.

Harris, Zellig S. 1968. *Mathematical Structures of Language*. New York: Wiley.

Holland, Richard C. G, Thomas A Down, Matthew R Pocock, Andreas Prlic, David Huen, Keith James, Sylvain Foisy, Andreas Dräger, Andy Yates, Michael Heuer & Mark J Schreiber. 2008. BioJava: An open-source frame-work for bioinformatics. *Bioinformatics* 24(18). 2096–2097.

Holton, Gary. this volume. Language archives: They're not just for linguists any more.

[5] http://www.lat-mpi.eu/tools/elan/

[6] http://www.lat-mpi.eu/tools/lexus/

Kemps-Snijders, Marc, Mark-Jan Nederhof & Peter Wittenburg. 2006. LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, 1862–1865.

Klamer, Marian. this volume. Tours of the past through the present of eastern Indonesia.

Needleman, Saul B & Christian D Wunsch. 1970. 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.

Quasthoff, Uwe & Christian Wolff. 2002. The Poisson collocation measure and its applications. In *Proceedings of the Second International Workshop on Computational Approaches to Collocations*, Vienna.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14(5). 465–471.

Senft, Gunter. 1983–1997. Tales of the Trobriand Islanders. Transcribed, morpheme-interlinearized and glossed corpus of Kilivila (fairy-)tales. Nijmegen: Mimeo.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Amit Kirschenbaum
amit@informatik.uni-leipzig.de

Peter Wittenburg
Peter.Wittenburg@mpi.nl

Gerhard Heyer
heyer@informatik.uni-leipzig.de