PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 105 APRIL 2016 63–76

An Algorithm for Morphological Segmentation of Esperanto Words

Theresa Guinard

New Mexico Institute of Mining and Technology

Abstract

Morphological analysis (finding the component morphemes of a word and tagging morphemes with part-of-speech information) is a useful preprocessing step in many natural language processing applications, especially for synthetic languages. Compound words from the constructed language Esperanto are formed by straightforward agglutination, but for many words, there is more than one possible sequence of component morphemes. However, one segmentation is usually more semantically probable than the others. This paper presents a modified n-gram Markov model that finds the most probable segmentation of any Esperanto word, where the model's states represent morpheme part-of-speech and semantic classes. The overall segmentation accuracy was over 98% for a set of presegmented dictionary words.

1. Introduction

Esperanto, a planned language developed in 1887, is purely agglutinative; compound words are formed by juxtaposing morphemes, where the spelling and pronunciation of the morphemes do not change during this process. The official rules for word formation are permissive, but in practice, producing an understandable compound word relies on complex semantic relationships between the morphemes.

Sometimes, an Esperanto word is morphologically ambiguous: there is more than one grammatically legal sequence of component morphemes. For example, the word "katokulo" can be segmented as "kat'okul'o", meaning "cat eye", as "kat'o'kul'o", meaning "cat-like gnat", or as "kat'ok'ul'o", which is grammatically permissible (by official rules), but has no discernible meaning. Usually, one segmentation is more semantically probable than the others.

© 2016 PBML. Distributed under CC BY-NC-ND. Corresponding author: tguinard@gmail.com
Cite as: Theresa Guinard. An Algorithm for Morphological Segmentation of Esperanto Words. The Prague Bulletin of Mathematical Linguistics No. 105, 2016, pp. 63–76. doi: 10.1515/pralin-2016-0003.

This study confronts the problem of morphological analysis: segmenting a word into component morphemes, and tagging the morphemes with part-of-speech information. This study takes a supervised approach, so I assume that a lexicon of tagged morphemes is given. Because the process for forming compound words is purely agglutinative, one can easily find the set of all possible segmentations for a given Esperanto word, but the main challenge is disambiguation.

Morphological analysis can potentially benefit a wide range of natural language processing applications, as individual word structures and meanings become easier to systematically interpret. For highly agglutinative or highly inflectional language, this is especially useful. In particular, for such languages, morphological analysis has been successfully applied to spell checking algorithms (Agirre et al., 1992) and (Solak and Oflazer, 1992), and machine translation (Lee, 2004) and (Goldwater and McClosky, 2005).

2. Overview of Esperanto Morphology

Esperanto morphemes can be categorized into four general categories: word endings, roots, affixes, and standalone words.

Word endings mark the part of speech of most words as a noun, verb, adjective, or adverb. Word endings also incorporate inflectional information. The morpheme "j" indicates whether a noun or adjective is plural; the word ending for a noun is "o", but the word ending for a plural noun is "oj". The morpheme "n" can be added to a noun, adjective, or adverb ending to mark the accusative case; "on" would signify an accusative noun, and "ojn" would signify a plural accusative noun. The accusative marker can also be appended to pronouns, and the plural and accusative markers can be appended to some correlatives. There are exactly six word endings for verbs, which indicate different tenses and moods: "i", "os", "as", "is", "u", and "us" respectively correspond to the infinitive, future tense, present tense, past tense, imperative, and conditional forms of the verb.

Roots make up the majority of Esperanto morphemes. A root has no definite part of speech, so in principle, any root can be combined with any word ending. For example, the root "pluv" is often used as a noun: "pluvo" ("rain"). However, "pluvi" (verb; "to rain"), "pluva" (adjective; "rain-like"), and "pluve" (adverb; "like rain") are all permissible Esperanto words. Although any word ending can be used, Kalocsay and Waringhien (1985) proposed that each root has an inherent part of speech. Currently, the official morpheme list provided by Akademio de Esperanto (2008) implements this idea, listing each root with the most frequent word ending.

Affixes can legally function in the same way as roots, but are usually prepended or appended to roots. For example, the prefix "mal" ("opposite") negates the meaning of the word it prepends: "bona" ("good") becomes "malbona" ("bad"), but "mal" can also function as an ordinary root: "malo" (noun; "opposite"). Similarly the suffix

"an" ("member") usually modifies a root: "klubo" ("club") becomes "klubano" ("club member"), but it can also form the word "ano" (noun; "member").

There is a notable class of suffixes, which are not used as roots in practice, but form participles to create compound tenses. One such suffix is "it", which can be appended to the verb root "skrib" ("to write") to form the phrase "estas skribita" ("has been written"). The suffixes in this class refer to different tenses ("has been written" vs. "is being written") and may refer to either the subject or object of the verb ("has been written" vs. "has been writing").

Standalone words are commonly-used words, including numbers, prepositions, pronouns, articles, exclamations, correlatives, and some adverbs. Correlatives are a class of function words including interrogatives ("what", "which"), demonstratives ("somehow", "somebody"), universals ("always", "everything"), and negatives ("nothing", "nobody"). Standalone morphemes most often appear uncompounded, but most can also act as component morphemes, whether this is through compounding with roots and other standalone morpheme, or adding a word ending. An example of standalone compounding is the word "dudekjara" ("twenty-year"), which contains the standalone morphemes "du" ("two") and "dek" ("ten"), the root "jar" ("year"), and the word ending "a" (adjective). The word "adiaŭi" ("to say goodbye") is formed using the standalone morpheme "adiaŭ" ("goodbye") and the word ending "i" (infinitive verb).

Forming compound words is a relatively permissive process. Fundamento de Esperanto, the official guide to Esperanto grammar, specifies only the basic mechanism for compound word formation (Zamenhof, 1905). Compound words are always formed by morpheme juxtaposition, and the principle morpheme occurs at the end of a word. For example, "ŝipvaporo" means "steam from a ship", while "vaporŝipo" means "steamship" (both words contain the morphemes "vapor" ("steam") and "ŝip" ("ship")). Roots can either be directly juxtaposed or separated by a word ending ("vaporŝipo" and "vaporoŝipo" are equivalent in meaning). The most common word ending to occur in the middle of words is "o", but the uninflected word endings "a", "i", "e" also often occur, as well as the accusative adverb ending "en". A word must always end with a word ending or a standalone morpheme, never with a root.

3. Previous Work

3.1. Other Agglutinative Languages

Koskenniemi (1984) proposed an influential morphological analysis model, the so-called "two-level morphology", which is applicable to languages with various morphologies, including agglutinative. The model consists of two components: a lexicon and a set of rules. The lexicon is a predefined list of tagged morphemes, and the rules

are a set of finite state transducers, which directly transform an input word into a list of tagged component morphemes.

The ideas used by Koskenniemi (using a set of categorized morphemes and representing morphological rules as a finite state model) have proved to be a useful starting point for many subsequent studies. Alegria et al. (1996) developed a morphological analysis pipeline for Basque, directly incorporating Koskenniemi's model. Other studies have incorporated statistical finite-state models, such as Markov models or conditional random fields, for disambiguation. Rios and Mamani (2014) implemented a morphological analysis system for the Quechua language, using finite state transducers to recognize possible morphological analyses, and conditional random fields to perform disambiguation. Hakkani-Tür et al. (2002) performed Turkish morphological disambiguation using hidden Markov models.

Depending on language-specific considerations, it is potentially useful to incorporate rule-based analysis steps that do not necessarily fit a finite-state model. Ezeiza et al. (1998) used a combination of constraint grammar rules and a hidden Markov model to disambiguate morpheme part-of-speech tags in Basque words. Nongmeikapam et al. (2012) performed morphological segmentation for Manipuri, incorporating Manipuri syllabification rules and an n-gram Markov model. Solak and Oflazer (1992) implemented a spelling checking system for Turkish using various phonological and morphological rules. The first segmentation found (via maximal morpheme matching) that follows these rules is accepted.

Like many of these previous approaches, I apply Koskenniemi's general approach to Esperanto. Morphemes are classified by part-of-speech and semantic properties, and an n-gram Markov model is used for disambiguation.

3.2. Esperanto

Some morphological analysis methods have been developed for Esperanto, but this is still a largely unexplored topic.

McBurnett (1985) wrote a morphological segmentation algorithm, which maximizes the lengths of morphemes as a word is scanned from left to right, incorporating a few rules to ensure a grammatically legal segmentation is found. For example, the accusative and plural markers must occur in a specific order after a word ending morpheme, and a word cannot end with a root or affix. Maximal morpheme matching has been incorporated into morphological analysis systems for other agglutinative languages, including German (for compound nouns only) (Lezius et al., 1998) and Turkish (Solak and Oflazer, 1992). Thus, it is valuable to directly compare McBurnett's approach to other approaches of Esperanto morphological segmentation.

Hana (1998) developed a two-level morphology system for Esperanto by descriptively analyzing word formation patterns. This system was able to recognize most Esperanto words in a corpus, and reported 13.6% morphological ambiguity.

Some Esperanto spell checkers use morphological considerations. Esperantilo is an application that contains a spell checker along with many other linguistic tools for Esperanto (Trzewik, 2006). The spell checker uses a list of base morphemes, each with a set of prefixes, suffixes, and word endings that are often used with the base morpheme. A word is evaluated using rule-based criteria, which ultimately limits the complexity of a word relative to known derivations. Blahuš (2009) used the Hunspell framework to write a spell checker for the open source word processor OpenOffice. This spell checker was implemented using pattern matching based on known, fixed-length morpheme combinations, where morphemes were categorized by semantic and part-of-speech properties. Although both of these systems work well for many words, neither fully encapsulates the agglutinative nature of Esperanto morphology.

This study attempts to construct an algorithm that can segment any Esperanto word, which requires the ability to process words with any number of morphemes. McBurnett's and Hana's approaches are directly applicable to this goal, though this study focuses on developing a statistical approach. I do experiment with adding a simple rule-based step, where some non-grammatical segmentations are discarded before disambiguation, though this is much less sophisticated than Hana's system.

4. Methods

This approach[1] focuses on using a modified n-gram Markov model for disambiguation, where states represent semantic and part-of-speech classes of morphemes. Various orders of n-gram Markov models were tried, as it is not immediately evident which value of n would be optimal.

In addition, I implemented a maximal morpheme matching algorithm, which uses a simple rule-based step that discards ungrammatical segmentations before disambiguation, similar to McBurnett's approach.

To evaluate the results of each disambiguation method, I compare the segmentation accuracy to the expected accuracy if a random valid segmentation is chosen.

For all outputs, only segmentation accuracy is reported, as opposed to tagging accuracy, as only a set of presegmented words was readily available. However, this is not a huge disadvantage, since most morphemes only belong to one class, as defined in this study.

Additionally, this method does not attempt to perform hierarchical disambiguation of morphological structure, e.g. determining whether to interpret "unlockable" as "[un+lock]able" ("able to unlock"), or as "un[lock+able]" ("not able to lock"). A hierarchical disambiguation step can be applied independently after segmentation, and for many applications, assuming a linear morphological structure may be sufficient.

[1] Source code available at `https://github.com/tguinard/EsperantoWordSegmenter`

General Category	Tags
Standalone	Adverb, Article, Conjunction, Correlative, Exclamation, Number, Preposition, Pronoun
Affix	AdjectiveSuffix, NounSuffix, NumberSuffix, PeopleAnimalSuffix, TenseSuffix, VerbSuffix, NounPrefix, PeopleAnimalPrefix, PrepositionPrefix, VerbPrefix
Root	Adjective, Adverb, Noun, PeopleAnimal, Verb
Mid-Word Endings	O, OtherMidWordEndings
Word Endings	AdjectiveEnding, AdverbEnding, NounEnding, PronounCorrelativeEnding, VerbEnding

Table 1. Morpheme Categorization

4.1. Datasets

4.1.1. Lexicon

All roots that occur in Esperantilo (Trzewik, 2006) were used, as well as all standalone and affix morphemes from Akademio de Esperanto (2008).

Akademio de Esperanto lists prefixes, suffixes, and standalone morphemes separately. I manually categorized standalone morphemes based on part of speech. Prefixes and suffixes were manually categorized by which kind of morphemes they often modify, barring two exceptions. Tense suffixes, used to create participles in compound tenses, were differentiated from verb suffixes. The preposition prefix class consists of morphemes that can act as either prepositions or prefixes.

Roots were categorized by part of speech, using the associated word endings provided by Esperantilo. I used one additional semantic class for roots: people and animals. I defined this class as any noun morpheme that can use the suffix “in” (which makes a word feminine). These morphemes were removed from the noun category.

Word endings were categorized manually by part of speech and whether the morpheme can be used in the middle of a word. Although the plural and accusative markers (“j” and “n”) are considered separate morphemes, all possible combinations of word endings, the plural marker, and the accusative marker were explicitly listed. For example, “o”, “oj”, “on”, and “ojn” were all listed as separate morphemes. However, the plural and accusative markers are also listed separately since they may modify pronouns and correlatives; the Markov model training set should only list the plural and accusative markers as separate morphemes in this case.

An overview of the tags used can be found in Table 1.

4.1.2. Training and Testing Sets: Presegmented Dictionary Words

The ESPSOF project lists over 50,000 Esperanto words segmented into component morphemes (Witkam, 2008). The word list was constructed from various Esperanto dictionaries, and the segmentations were manually adjusted by Witkam. Only a subset of this list is used as input for this study since not all of the roots used in ESPSOF are listed in Esperantilo.

The total size of this input set is 42,356 words, which were split into a training set and test set (respectively used to set and test the Markov model parameters). Three-quarters of the words were used in the training set, and one-quarter in the test set. This three-quarters split was held over words with a consistent number of morphemes (e.g. three-quarters of words with two morphemes are in the training set). For all experiments run in this study, the same test set and training set were used.

Setting the Markov model parameters requires these segmentations to be tagged. Most morphemes belong to only one class as defined in this study, but for those that belong to multiple classes, simple rules are applied to determine the correct tag. For example, roots and word endings should match in part of speech if possible. If there is still uncertainty in the correct tag to assign, all possible tags are used with equal weight, but the total influence of each word on the Markov model is equal.

4.2. Segmentation Algorithm with Markov Model

There are two steps to the segmentation algorithm: finding all possible segmentations using a trie lookup algorithm, then selecting the best segmentation using a Markov model.

4.2.1. Segmentation

The segmentation phase finds all morpheme sequences that form the input word when juxtaposed. During this step, a minimalistic set of rules may be optionally applied:

- A word cannot end with a root or affix.
- The accusative marker “n” and the plural marker “j” can only appear after pronouns or correlatives (or after some word endings, but this is built into the lexicon).
- The definite article “la” cannot be combined with other morphemes.

All morpheme sequences are found via trie lookup.

For the ESPSOF word list, when the rules are applied, a word has a mean of 2.15 segmentations, 53.5% of words have at least two possible segmentations, and the largest number of distinct segmentations is 112. Thus, disambiguation is necessary.

4.2.2. Disambiguation

Disambiguation is performed using a modified n-gram Markov model. Each state represents n morpheme classes.

For the unigram model, each traversal begins on a state called "Start", visits the states corresponding to each morpheme class, and finishes on a state called "End". For example, in the segmentation "kat'okul'o", the individual morphemes are Esperanto for "cat", "eye", and (noun ending). The sequence of states visited is:

$$Start \rightarrow PeopleAnimal \rightarrow Noun \rightarrow NounEnding \rightarrow End$$

The frequency of each transition in the training set is used to calculate probabilities used by the Markov model.

The probability that the current state is B, given that the previous state was A, or $P(B|A)$, is related to the frequency of transitions from A to B, or $|(A, B)|$, and the sum of the frequency of transitions from state A to any state, S, or $|(A, S)|$.

$$P(B|A) = \frac{|(A, B)|}{\sum_{S \in States} |(A, S)|}$$

The score of the traversal, T, is calculated as follows. $|new\_class(B)|$ is the number of morphemes represented the last morpheme class in state B's n-gram, and α is a positive real number. For each word, the segmentation with the highest score is accepted as the correct segmentation. Occasionally, more than one segmentation may share the highest score. If this is the case, the ambiguity is resolved via maximal morpheme matching.

$$score(T) = \prod_{(A,B) \in T} \frac{\alpha \cdot P(B|A)}{|new\_class(B)|}$$

If α is omitted, this forms a straightforward Markov model, adjusted for unequal morpheme class sizes. Including α changes how often longer morpheme sequences are preferred. An optimal value for α can be found empirically in the training set.

For the bigram Markov model, each state represents two consecutive tags, and for the trigram Markov model, each state represents three consecutive tags. The beginning state always represents n Start tags. For example, the transition sequence of "kat'okul'o" for the bigram model is:

$$(Start \cdot Start) \rightarrow (Start \cdot PeopleAnimal) \rightarrow (PeopleAnimal \cdot Noun) \rightarrow$$
$$(Noun \cdot NounEnding) \rightarrow (NounEnding \cdot End)$$

For all models, the score calculation is equivalent, including the value of α (the number of states is constant between models for a given segmentation).

4.3. Additional Tests

4.3.1. Maximal Morpheme Match

This algorithm uses the same segmentation phase as the Markov model approach, but then selects the segmentation where the initial morphemes are as long as possible. That is, the length of the first morpheme is maximized, and if there is still ambiguity, the length of the subsequent morpheme is maximized, and this is repeated until there is no ambiguity.

The performance of this algorithm was compared with the Markov models by running this algorithm on all words from the ESPSOF word list (i.e. both the training set and the test set).

4.3.2. Randomly Selecting a Segmentation

As a baseline for comparing accuracy, I calculated the expected accuracy of randomly selecting a segmentation after the initial segmentation phase. This was applied to all words from the ESPSOF word list.

5. Results

When evaluating segmentation accuracy, a segmentation is considered correct if it equivalent to the expected segmentation, with one exception: the output segmentation contains a morpheme that appears in Esperantilo but not in ESPSOF, and this morpheme can be constructed from multiple morphemes in the expected solution. By inspecting the output, this is caused by Esperantilo listing morphemes that could be considered the combination of several morphemes. As an example, ESPSOF segments “prezidanto” (“president”) as “prezid’ant’o” (“someone who presides”), while Esperantilo lists “prezidant” as a separate morpheme, so the output segmentation is “prezidant’o”.

5.1. Various n-gram Markov Models

The segmentation accuracies of the three Markov models, with no rule-based step, are shown in Table 2. Although accuracies of the Markov models are high overall, there is a definite decrease in accuracy as the number of morphemes per word increases. All three models perform very similarly, though the higher order n-gram models are slightly more accurate overall.

This approach implements maximal morpheme matching as a secondary line of disambiguation in the case that multiple segmentations share the same highest score (this happened about 0.2-0.3% of the time). Depending on the model and word set, this strategy correctly resolved between 61-76% of these ambiguities.

Number of Morphemes	1	2	3	4	5	6	7	Any
Percent of Input Words	0.378	30.1	47.3	19.3	2.81	0.168	0.0142	100
Unigram: Training Set	1.00	1.00	0.990	0.966	0.909	0.811	0.750	0.986
Unigram: Test Set	1.00	0.999	0.989	0.963	0.906	0.944	1.00	0.985
Bigram: Training Set	1.00	1.00	0.992	0.971	0.936	0.906	1.00	0.989
Bigram: Test Set	1.00	1.00	0.991	0.969	0.923	0.833	0.500	0.989
Trigram: Training Set	1.00	1.00	0.992	0.971	0.933	0.962	1.00	0.989
Trigram: Test Set	1.00	1.00	0.991	0.973	0.916	0.833	0.500	0.987

Table 2. Markov model segmentation accuracies (no rules applied)

In terms of the errors that did occur, I observed that some were due to the inconsistent segmentation technique present in the ESPSOF word list. For example, ESPSOF segments the correlative "nenio" ("nothing") as a root and word ending in "neni'o'far'ul'o" ("person who does nothing"), but other correlatives are treated as standalone morphemes, such as "nenies" ("nobody's") in "nenies'land'o" ("no man's land"). Additionally, ESPSOF segments "esperanto" as "esper'ant'o" ("one who hopes"), which is the original etymology of the language's name, but "esperant" is used as a single morpheme elsewhere in the list. These inconsistencies seem to account for approximately 10% of the total errors for each model.

In the test sets of each model, the erroneous segmentations had the same number of morphemes as the ESPSOF segmentation 57-61% of the time. The erroneous segmentations had too few morphemes 35-41% of the time and too many morphemes 2-4% of the time.

For the segmentations that had too few morphemes, most of the errors were common between all three models in the test set. 49 of these errors were common between all three models, while the unigram model had the most such errors (57). For all models, 72-76% of these erroneous segmentations combined two morphemes from ESPSOF's solution to form a single morpheme. For example, "hufofero" should be segmented as "huf'o'fer'o" ("horseshoe", literally "hoof iron"), but each model produced "huf'ofer'o", ("hoof offer"). This type of error seems tricky to overcome, espe-

cially when the merged morpheme has a similar semantic class to the two separated morphemes.

There were very few instances where a segmentation with too may morphemes was produced, but this occurred most often in the unigram model (6 errors, vs. 3 each for the bigram and trigram models). The extra errors for the unigram model were due to overfavoring the accusative "n" morpheme. For example, "vinmiksaĵo" should be segmented as "vin'miks'aĵ'o" ("wine mixture"), but the unigram model produced "vi'n'miks'aĵ'o" (nonsense, literally "you (accusative) mixture").

The majority of the variation between the three models came from instances where the segmentation produced had the same number of morphemes as expected. There were 100 such errors for the unigram model, 82 for the bigram, and 76 for the trigram. 50 of these errors were common between all three models. These errors most directly show where Esperanto morphology does not follow a specific n-gram model, as the α factor does not influence these errors. For example, the unigram model erroneously uses mid-word endings more often than the bigram and trigram models, e.g. "help'a'gad'o" ("helpful cod") instead of "help'ag'ad'o" ("acting helpful").

Some of the errors that were not caused by inconsistencies in ESPSOF's segmentation may be resolved by improving the tag set. The presented morpheme categorization was effective, but optimal categorization is still an open issue.

5.2. Comparison with Maximal Matching and Random Selection

Table 3 compares the unigram Markov model with the maximal morpheme matching algorithm and the random selection strategy.

In terms of overall accuracy, the Markov model is significantly more accurate than maximal matching, though both developed algorithms are significantly more accurate than randomly choosing a segmentation.

The accuracy of the random selection method notably decreases as the number of morphemes increases, so it is natural for any segmentation algorithm to perform worse as the number of morphemes per word increases.

The maximal matching's performance is much more sensitive to the number of morphemes per word than the Markov model is. For words with only two morphemes, maximal matching performs comparably to the Markov model, but the accuracy quickly drops as the number of morphemes increases, approaching the accuracy of the random selection method.

When adding the rule-based step to the Markov models, the performance only changed for the test set of the unigram and trigram models, which correctly segmented one and two additional words respectively. However, adding rules significantly improves the accuracy of the maximal matching and random selection methods, as seen in Table 3.

Number of Morphemes	1	2	3	4	5	6	7	Any
Percent of Input Words	0.378	30.1	47.3	19.3	2.81	0.168	0.0142	100
Unigram: Training Set	1.00	1.00	0.990	0.966	0.909	0.811	0.750	0.986
Unigram: Test Set	1.00	0.999	0.989	0.963	0.906	0.944	1.00	0.985
Maximal Matching	1.00	1.00	0.970	0.833	0.676	0.577	0.333	0.944
Maximal Matching: No Rules	1.00	0.995	0.948	0.801	0.638	0.535	0.333	0.925
Random Selection	0.902	0.709	0.685	0.623	0.538	0.428	0.412	0.676
Random Selection: No Rules	0.750	0.630	0.541	0.437	0.330	0.202	0.208	0.542

Table 3. Comparison of Markov model, maximal matching, and random selection segmentation accuracies

6. Conclusion

This study investigated an n-gram Markov model approach to Esperanto morphological segmentation, as well as a maximal matching approach for comparison. An extra factor was added to the Markov model to adjust how often longer sequences of morphemes are accepted. Morphemes were categorized by part of speech, with a few extra subclasses, which was sufficient for producing a high segmentation accuracy.

There was not much difference between the performances of the various n-gram orders, although the bigram and trigram models were slightly more accurate for both the training and test sets. Both the Markov model and maximal matching approaches performed significantly better than randomly selecting a valid dissection, but the Markov model is more scalable to words with more morphemes. The rule-based step used in this study was useful for improving the accuracy of the maximal matching algorithm, but had no significant impact on the Markov model performances.

Bibliography

Agirre, Eneko, Inaki Alegria, Xabier Arregi, Xabier Artola, A Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. In *Proceedings of the third conference on Applied natural language processing*, pages 119–125. Association for Computational Linguistics, 1992.

Akademio de Esperanto. Akademia Vortaro, 2008. URL `http://akademio-de-esperanto.org/akademia_vortaro/`.

Alegria, Iñaki, Xabier Artola, Kepa Sarasola, and Miriam Urkia. Automatic morphological analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203, 1996.

Blahuš, Marek. Morphology-Aware Spell-Checking Dictionary for Esperanto. *RASLAN 2009 Recent Advances in Slavonic Natural Language Processing*, page 3, 2009.

Ezeiza, Nerea, Iñaki Alegria, José María Arriola, Rubén Urizar, and Itziar Aduriz. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 380–384. Association for Computational Linguistics, 1998.

Goldwater, Sharon and David McClosky. Improving statistical MT through morphological analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 676–683. Association for Computational Linguistics, 2005.

Hakkani-Tür, Dilek Z, Kemal Oflazer, and Gökhan Tür. Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*, 36(4):381–410, 2002.

Hana, Jiří. Two-level morphology of Esperanto. Master's thesis, Charles University Prague, Faculty of Mathematics and Physics, 1998. URL `http://www.ling.ohio-state.edu/~hana/esr/thesis.pdf`.

Kalocsay, Kálmán and Gaston Waringhien. *Plena Analiza Gramatiko de Esperanto*, volume 2. Universala Esperanto-Asocio, 1985.

Koskenniemi, Kimmo. A general computational model for word-form recognition and production. In *Proceedings of the 10th international conference on Computational Linguistics*, pages 178–181. Association for Computational Linguistics, 1984.

Lee, Young-Suk. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60. Association for Computational Linguistics, 2004.

Lezius, Wolfgang, Reinhard Rapp, and Manfred Wettler. A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 743–748. Association for Computational Linguistics, 1998.

McBurnett, Neal. Ilaro por Esperantaj Redaktoroj (Ilaro - a Toolkit for Esperanto Editors), 1985. URL `http://bcn.boulder.co.us/~neal/`.

Nongmeikapam, Kishorjit, Vidya Raj, Rk Yumnam, and Nirmal Sivaji Bandyopadhyay. Manipuri Morpheme Identification. *24th International Conference on Computational Linguistics*, pages 95–107, 2012.

Rios, Annette and Richard Castro Mamani. Morphological Disambiguation and Text Normalization for Southern Quechua Varieties. *COLING 2014*, page 39, 2014.

Solak, Aysin and Kemal Oflazer. Parsing agglutinative word structures and its application to spelling checking for Turkish. In *Proceedings of the 14th conference on Computational linguistics-Volume 1*, pages 39–45. Association for Computational Linguistics, 1992.

Trzewik, Artur. Esperantilo - text editor with particular Esperanto functions, spell and grammar checking and machine translation, 2006. URL `http://www.xdobry.de/esperantoedit/index_en.html`.

Witkam, Toon. ESPSOF (Esperanto-Softvaro), 2008. URL `http://www.espsof.com/`.

Zamenhof, L. L. Fundamento de Esperanto, 1905. URL `http://akademio-de-esperanto.org/fundamento/gramatiko_angla.html`.

Address for correspondence:
Theresa Guinard
`tguinard@gmail.com`
16801 NE 39th Ct #F2020
Redmond, WA, USA 98052