# Using Resource-Rich Languages to Improve Morphological Analysis of Under-Resourced Languages

**Peter Baumann, Janet Pierrehumbert**

Northwestern University, Evanston, IL, USA
baumann@u.northwestern.edu, jbp@northwestern.edu

## Abstract

The world-wide proliferation of digital communications has created the need for language and speech processing systems for under-resourced languages. Developing such systems is challenging if only small data sets are available, and the problem is exacerbated for languages with highly productive morphology. However, many under-resourced languages are spoken in multi-lingual environments together with at least one resource-rich language and thus have numerous borrowings from resource-rich languages. Based on this insight, we argue that readily available resources from resource-rich languages can be used to bootstrap the morphological analyses of under-resourced languages with complex and productive morphological systems. In a case study of two such languages, Tagalog and Zulu, we show that an easily obtainable English wordlist can be deployed to seed a morphological analysis algorithm from a small training set of conversational transcripts. Our method achieves a precision of 100% and identifies 28 and 66 of the most productive affixes in Tagalog and Zulu, respectively.

**Keywords:** morphology, language contact, code switching

## 1. Introduction

Globalization and the rise of the internet as a global medium of communication has led to an ever increasing demand for natural-language processing (NLP) systems. However, the number of languages for which such systems are available is small compared to the nearly 7000 languages (Maxwell and Hughes, 2006) spoken on the planet. Most languages are under-resourced, lacking not only practical NLP systems, but even the large labeled corpora typically used to develop such systems. As globalization continues, some of these under-resourced languages may suddenly acquire imminent political or economic interest, so that rapid development of NLP systems for them is needed. There is growing interest in the development of NLP or speech-processing systems for under-resourced languages (Le and Besacier, 2009), but such attempts are often limited by difficulties in collecting the necessary datasets (Lewis and Yang, 2012).

The starting point for our study is the observation that many under-resourced languages share a feature that can reduce the amount of data needed and provide a short-cut in building an NLP system: They are spoken in multi-lingual environments together with at least one resource-rich language. Code-switching and borrowing from this resource-rich language is common. Examples of such interactions between under-resourced and resource-rich languages include Quechua and Spanish in the Andes (Sanchéz, 2003), Azeri and Russian in Azerbaijan (Zuercher, 2010), Tagalog and English in the Philippines (Bautista and Bolton, 2008), and Zulu and English in South Africa (Ramsay-Brijball, 1999). For a comprehensive overview of loanwords and borrowings in many under-resourced languages, see (Haspelmath and Tadmor, 2009)[1].

This paper is a case study in the use of readily available resources from a resource-rich language (English) to bootstrap the morphological analyses of two under-resourced language with complex and productive morphological systems (Tagalog and Zulu). We demonstrate that a list of words found in English movie subtitles can be used to extract *linguistically accurate* sets of prefixes and suffixes from small corpora of conversational Tagalog and Zulu.

We target precision as the measure of interest because in many applications of computational morphology, precision is more important than recall, and can be estimated more reliably given only limited resources: For retrieval of text and audio documents, precision can estimated from user satisfaction with search results, but large amounts of unlabeled material can make it impossible to tabulate misses (Brin and Page, 1998). Furthermore, false alarms in morphological decomposition have adverse consequences if they create spurious word associations in systems that employ stemming or lemmatization. More generally, affix frequencies obey a Zipfian distribution. Correctly identifying the most productive affixes is thus more important for the analysis of out-of-vocabulary (OOV) words in any unrestricted system, whereas missing the numerous marginally productive affixes that exist in many languages can have little importance.

We argue that using resources from resource-rich languages to identify morphological features of under-resourced languages allows for a faster deployment of NLP or speech processing systems than recruiting informants to generate a labeled training set, especially since for many under-resourced languages reliable informants may not be readily available. And even if reliable informants are available, our approach provides a fast method to extract the most common morphological features, which can save valuable time for informants to provide more fine-grained morphological analyses.

[1] Online access to the authors' database is available at `http://wold.livingsources.org/`.

## 2. Morphological Analysis and Related Work

For morphologically impoverished languages like English, words may serve as the most practical approximation to morphemes (Church, 2005). However, this is not feasible for morphologically rich languages like Tagalog or Zulu, in which the productive recombination of morphemes means that vocabularies derived from training materials offer poor coverage of new materials. Given observed vocabularies in the range of 50,000 to 70,000 words, Kurimo et al. (2006) report OOV rates in previously unseen materials that are an order of magnitude higher for Finnish, Estonian, and Turkish than for English (see also section 3.2.).

When developing systems for languages with rich morphology from limited resources, morphological analysis becomes indispensable. In addition, morphological information has proven useful in information retrieval (Schulz et al., 2002), in determining semantic relationships between words (Namer and Zweigenbaum, 2004) or in improving language models for large-vocabulary continuous-speech recognition (El-Desoky et al., 2009; El-Desoky Mousa et al., 2012) and machine translation (El Kholy and Habash, 2012).

### 2.1. Unsupervised Morphological Analysis

While classic approaches to morphological analysis require labelled training sets, recent work in morphological analysis has been focused on unsupervised methods, which do not require any expert knowledge or labeled data (Goldsmith, 2001; Monson et al., 2004; Bernhard, 2006; Dasgupta and Ng, 2007). These algorithms try to identify recurrent string patterns to derive a morpheme lexicon, which is optimal in some sense. One such system is *Morfessor* (Creutz and Lagus, 2007), which is based on the minimum-description-length principle, and tries to find a lexicon of morphemes that is both accurate and minimal, and captures concatenative morphological processes of affixation and compounding using a hidden-Markov model over sequences of prefixes, stems and suffixes. While Morfessor derives its lexicon of morphemes in one single optimization procedure, other systems (Bernhard, 2006; Dasgupta and Ng, 2007) follow a multi-stage approach, in which prefixes and suffixes are identified in a first stage. This set of affixes is then used to identify potential stems, which are in turn used to obtain a segmentation of all words in the corpus. Despite their conceptual differences, both approaches were implemented in systems that were both among the top performers in the MorphoChallenge (Kurimo et al., 2009), a competition of unsupervised algorithms for morphological learning.

Our approach to morphological analysis is related to those described in (Bernhard, 2006; Dasgupta and Ng, 2007), as it is a multi-stage system that begins by finding affixes. However, we do not identify potential affixes in a fully unsupervised manner, but use one additional resource: an English word list. Through this multi-lingual approach we expect to achieve higher precision than fully unsupervised systems can normally attain. However, it should be emphasized that our method does not require any labeled data, neither in the target languages (Tagalog and Zulu) nor in English, so that it is much closer to an unsupervised method than to a semi-supervised one.

### 2.2. Multilingual Morphological Analysis

Multilingual resources can significantly improve morphological analysis in under-resourced languages by using large labelled data in one language and a method to project labels from one language to another. This method can be applied to learn Arabic morphology based on an English stemmer and a small parallel training corpus (Rogati et al., 2003). For two languages within the same language family, it is even possible to dispense with the parallel training set and exploit similar distributional patterns across relate language to obtain morpho-syntactic annotations in one language based on labelled data in another (Hana et al., 2004; Feldman and Hana, 2010). Finally, parallel data sources have been shown to improve the performance of unsupervised methods in both languages, as in the case of parallel morphological analysis of Arabic and Hebrew (Snyder and Barzilay, 2008).

While all these approaches are similar in spirit to the method reported here, they crucially rely on either labelled data sets in one language or (at least small) parallel corpora between two languages. Our approach does require such structured data. Instead we will show that a simple English word list can be used to obtain a morphological analysis of two under-resourced languages (Tagalog and Zulu), requiring only small data sets in those languages.

## 3. Case Study

### 3.1. Languages

Tagalog (Filipino), an Austronesian language, is the national language of the Philippines. It has rich morphology which is mainly expressed through prefixes, but it also shows suffixation, infixation, and reduplication (Schachter and Otanes, 1983). Due to the colonial history of the Philippines, Tagalog has a large number of loan words from Spanish and more recently from English. These words participate in the morphological system of Tagalog (Bautista and Bolton, 2008). While the Spanish loan words have undergone spelling changes, loan words from English retain their English spelling, as illustrated in (1).

(1) a. pinag + swimming + an
prefix + stem + suffix
b. nag + feedback
prefix + stem
c. interview + hin
stem + suffix

Zulu (isiZulu), a Bantu language, is one of the major languages of South Africa. It is an agglutinating language with a rich inventory of prefixes and suffixes. Like Tagalog, Zulu has a large number of English loan words, which participate in the morphological system of Zulu, but have retained their English spelling, as illustrated in (2)

(2) a. uku + understand + a
prefix + stem + suffix
b. uya + complain + er
prefix + stem + suffix (English)

The fact that the most recent borrowings are not yet integrated into the orthographical system of the host language is not restricted to Tagalog and Zulu, but a rather common phenomenon, which can even be observed in languages with highly standardized orthographies like German or French, where English loan words retain their original spelling. This particular situation makes it feasible to directly use a list of English words to identify Tagalog or Zulu affixes.

### 3.2. Data

The data for our experiments are simple word lists. The lists of Tagalog and Zulu words were compiled from transcripts of 80 hours of Tagalog and Zulu telephone dialogues, respectively, provided by the IARPA Babel Program (Full Language Packs of Tagalog language collection release *IARPA-babel106-v0.2g* and Zulu language collection release *IARPA-babel206b-v0.1e*; see Acknowledgments). The transcripts have the size of around $522,000$ word tokens for Tagalog and around $345,000$ for Zulu, and the final word list used in our experiment contain $16,655$ unique word types for Tagalog and $54,009$ word types for Zulu.

As an illustration of the differences in vocabulary growth between Zulu, Tagalog and English, Figure 1 shows the number of word types for different numbers of word tokens. The English data is taken from Switchboard (Godfrey and Holliman, 1997).

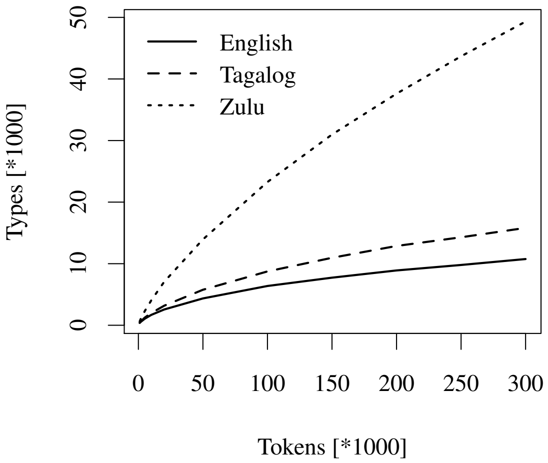

Figure 1: Type counts over token counts for Zulu, Tagalog, and English (Switchboard)

The English word list was compiled from `opensubtitles.org`[2]. This list was used because movie dialogues provide a sample of informal language. The word list contains all English words which occurred more than 30 times in movie subtitles. From this list, we excluded all words shorter than four characters, leaving a total of $46,862$ word types.

### 3.3. Method

Like many other methods (Monson et al., 2004; Bernhard, 2006; Dasgupta and Ng, 2007) in computational morphology, our algorithm operates at the level of word types and takes an English word list (EL) and word list of the target language (TL, Tagalog or Zulu) as inputs.

Since English loan words in participate in the Tagalog/Zulu morphological system, elements that precede or follow an identifiable English word are potential prefixes or suffixes, and so in a first step potential affixes are identified by searching for English words as substrings of words in TL. To reduce the likelihood of finding English affixes and to improve performance, both word lists are sorted by length in decreasing order.

However, some matches between English and Tagalog or Zulu matches are coincidental, and so in a second step all potential affixes, which occurred more than once, are validated against all TL words, for which the first step did not yield a hit. We assume that an affix is only valid if stripping it off a word yields other valid word forms (potentially stems) and if some number $w$ of these word forms can be found in the corpus. The ratio of $w$ over the total number of occurrences $a$ of a given affix in the corpus is logarithmically proportional to the mutual information between a potential affix and the set of attested word forms containing that affix, thus yielding a measure of association strength between affixes and attested word forms (Church and Hanks, 1990).

Since languages vary in the extent to which stripping an affix yields a valid word, a threshold for $\frac{w}{a}$ can only be determined in relation to the distribution of $\frac{w}{a}$ for all the affix candidates (Kilgarriff, 2009; Kilgarriff, 2005). For both languages, the distribution of $\frac{w}{a}$ was bimodal, with a valley of frequency zero falling between the two modes. Setting a decision threshold at this location provided a straightforward way to separate valid affixes, which have high mutual information with attested forms, from invalid affixes. As an example, Figure 2 shows the distribution of the ratio $\frac{w}{a}$ for Tagalog prefixes and the decision threshold.

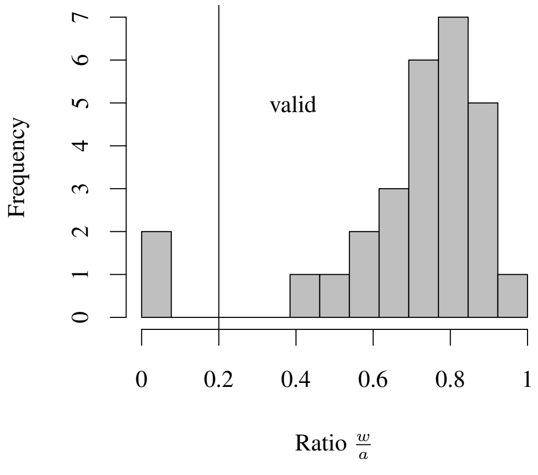

Figure 2: Distribution of the ratio $\frac{w}{a}$ for Tagalog prefixes. The vertical line marks the decision boundary.

The following pseudo-code presents a summary of our

[2] Available at `http://invokeit.wordpress.com/frequency-word-lists/`.

method:

```
0. Order EL and TL by length (in decreasing order)

1. Find potential affixes:
     for all words in TL do
       • if word contains any element from EL as a substring:
           – split word around substring
           – add prefix to list of potential prefixes
           – add suffix to list of potential suffixes
           – remove word from TL

2. Validate potential affixes:
     for all potential prefixes and suffixes do
       • $w \leftarrow$ count how often prefix (suffix) occurs as a prefix (suffix) in TL and resulting stem is a word in TL
       • $a \leftarrow$ count how often prefix (suffix) occurs as a prefix (suffix) in TL
```

### 3.4. Results

#### 3.4.1. Tagalog

For Tagalog, our method identified four suffixes (4) and 24 prefixes (3). We checked the affixes against a Tagalog reference grammar (Schachter and Otanes, 1983) and found that all identified prefixes can appear as Tagalog prefixes: the 20 prefixes in (3-a) are inflectional prefixes and serve to express different forms of grammatical aspect and focus; the single prefix in (3-b) is a derivational prefix to express politeness; the prefixes in (3-c) are the infix *-in-*, which on words starting with a vowel it occurs as a prefix, and the two prefixes *pa-* and *pag-* with the infix *-in-* inserted.

(3) a. ma-, mag-, na-, nag-, ka-, maka-, naka-, magka-, nagka-, i-, mai-, pa-, magpa-, nagpa-, ipa-, pag-, makapag-, nakapag-, pang-, ni-
b. paki-
c. in-, pina-, pinag-

Among the suffixes, the first three are actual Tagalog suffixes, while *-s* is the most common suffix in English, which is frequently used in the English sub-vocabulary of Tagalog.

(4) -an, -in, -hin, -s

#### 3.4.2. Zulu

For Zulu, our method identified nine suffixes (6) and 57 prefixes (5). We checked the affixes against a (small) annotated corpus of Zulu (Spiegler et al., 2010) and found that all identified prefixes are actual prefixes or combinations of prefixes in Zulu.

(5) a-, aba-, ama-, ba-, be-, bengi-, e-, eyi-, i-, ine-, iya-, iyi-, ka-, ko-, ku-, kule-, kuma-, kune-, kwa-, kwakuyi-, kwe-, kwi-, lama-, le-, ma-, nama-, ne-, nga-, ngama-, ngase-, nge-, ngi-, ngiku-, ngise-, ngiya-, ngo-, ni-, no-, o-, s-, se-, si-, u-, ube-, uku-, ukuyi-, una-, une-, use-, uya-, uyo-, uzo-, wa-, we-, ye-, yi-, zi-

Among the suffixes, there were six actual Zulu suffixes and the three English suffixes *-s*, *-er* and *-r*.

(6) -i, -e, -ile, -ayo, -o, -eka, -s, -er, -r

Crucially, not a single unattested affix was postulated for either language, and so our method achieves 100% precision on both prefixes and suffixes in both languages.

## 4. Discussion

The method presented in this paper is a proof of concept, showing that a readily available word list of a resource-rich language (English) can be used to identify a set of morphological features of two under-resourced languages (Tagalog and Zulu) with high precision. The method may not be strictly unsupervised, because it uses an English wordlist in addition to the target data sets. However, unlike (semi-)supervised or multilingual methods, it does not require any labelled data in either language. Given that our method was successfully applied to two unrelated languages, the overall approach holds promise for the analysis of other under-resourced languages with substantial borrowings from resource-rich languages. In addition to offering high precision, the method is also notable for its success using only a small training set.

## 5. Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0014. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

## 6. References

Bautista, M. A. Lourdes S. and Bolton, Kingsley. (2008). *Philippine English: Linguistic and Literary*. Hong Kong University Press, Hong Kong.

Bernhard, Delphine. (2006). Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of the PASCAL Challenges Workshop on Unsupervised Segmentation of Words into Morphemes*, pages 19–23, Venezia, Italy.

Brin, Sergey and Page, Lawrence. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117.

Church, Kenneth Ward and Hanks, Patrick. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Church, Kenneth W. (2005). The DDI Approach to Morphology. In Arppe, Antti, Carlson, Lauri, Lindén, Krister, Piitulainen, Jussi, Suominen, Mickael, Vainio, Martti, Westerlund, Hanna, and Yli-Jyrä, Anssi, editors, *Inquiries into Words, Constraints, and Contexts*, pages 25–34. CLSI Publications, Stanford.

Creutz, Mathias and Lagus, Krista. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34, February.

Dasgupta, Sajib and Ng, Vincent. (2007). High-performance, language-independent morphological segmentation. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 155–163.

El-Desoky, Amr, Gollan, Christian, Rybach, David, Schlüter, Ralf, and Ney, Hermann. (2009). Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR. In *Proceedings of Interspeech*, pages 2679–2682, Brighton, UK.

El-Desoky Mousa, Amr, Basha Shaik, M. Ali, Schlüter, Ralf, and Ney, Hermann. (2012). Morpheme level feature-based language models for German LVCSR. In *Proccedings of Interspeech*, Portland, OR, USA, September.

El Kholy, Ahmed and Habash, Nizar. (2012). Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26(1):25–45.

Feldman, Anna and Hana, Jirka. (2010). *A Resource-Light Approach to Morpho-Syntactic Tagging*, volume 70 of *Language and Computers: Studies in Practical Linguistics*. Rodopi.

Godfrey, John J. and Holliman, Edward. (1997). *Switchboard-1 Release 2*. Linguistic Data Consortium, Philadelphia.

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

Hana, Jirka, Feldman, Anna, and Brew, Chris. (2004). A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing( EMNLP)*, volume 4.

Haspelmath, Martin and Tadmor, Uri, editors. (2009). *Loanwords in the World's Languages: A Comparative Handbook*. Mouton de Gruyter, Berlin.

Kilgarriff, Adam. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2):263–276.

Kilgarriff, Adam. (2009). Simple maths for keywords. In Mahlberg, Michaela, González-Díaz, Victorina, and Smith, Catherine, editors, *Proceedings of the Corpus Linguistics Conference (CL2009)*, Liverpool, July.

Kurimo, Mikko, Puurula, Antti, Arisoy, Ebru, Siivola, Vesa, Hirsimäki, Teemu, Pylkkönen, Janne, Alumäe, Tanel, and Saraclar, Murat. (2006). Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2006)*, pages 487–494.

Kurimo, Mikko, Virpioja, Sami, Turunen, Ville T., Blackwood, Graeme W., and Byrne, William. (2009). Overview and results of morpho challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece.

Le, Viet-Bac and Besacier, Laurent. (2009). Automatic speech recognition for under-resourced languages: Application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1471–1482.

Lewis, William D. and Yang, Phong. (2012). Building MT for a severely under-resourced language: White Hmong. In *Association for Machine Translation in the Americas*, October.

Maxwell, Mike and Hughes, Baden. (2006). Frontiers in linguistic annotation for lower-density languages. In *Proceedings of COLING/ACL2006 Workshop on Frontiers in Linguistically Annotated Corpora*.

Monson, Christian, Lavie, Alon, Carbonell, Jaime, and Levin, Lori. (2004). Unsupervised induction of natural language morphology inflection classes. In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 52–61, Barcelona, Spain.

Namer, Fiammetta and Zweigenbaum, Pierre. (2004). Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Proceedings of Medinfo*, volume 11, pages 535–539, San Francisco, CA.

Ramsay-Brijball, Malini. (1999). Understanding Zulu-English code-switching: A psycho-social perspective. *South African Journal of Linguistics*, 17(2-3):161–172.

Rogati, Monica, McCarley, Scott, and Yang, Yiming. (2003). Unsupervised learning of Arabic stemming using a parallel corpus. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 391–398. Association for Computational Linguistics.

Sanchéz, Liliana. (2003). *Quechua-Spanish Bilingualism: Interference and Convergence in Functional Categories*. John Benjamins.

Schachter, Paul and Otanes, Fe T. (1983). *Tagalog Reference Grammar*. University of California Press, Berkeley.

Schulz, Stefan, Honeck, Martin, and Hahn, Udo. (2002). Biomedical text retrieval in languages with a complex morphology. In *CL Workshop on Natural Language Processing in the Biomedical Domain*, pages 61–68, Philadelphia, PA.

Snyder, Benjamin and Barzilay, Regina. (2008). Unsupervised multilingual learning for morphological segmentation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 737–745.

Spiegler, Sebastian, van der Spuy, Andrew, and Flach, Peter A. (2010). Ukwabelana - an open-source morphological Zulu corpus. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.

Zuercher, Kenneth. (2010). *Azerbaijani-Russian Code-switching And Code-mixing: Form, Function, And Identity*. Ph.D. thesis, University of Texas, Arlington, TX.