Morphological parsing of Swahili using crowdsourced lexical resources

Patrick Littell, Kaitlyn Price, Lori Levin
University of British Columbia, Oberlin College, Carnegie Mellon University
Vancouver, BC V6T1Z4, Oberlin, OH 44074, Pittsburgh, PA 15213
littell@interchange.ubc.ca, kprice@cs.oberlin.edu, lsl@cs.cmu.edu

Abstract

We describe a morphological analyzer for the Swahili language, written in an extension of XFST/LEXC intended for the easy declaration of morphophonological patterns and importation of lexical resources. Our analyzer was supplemented extensively with data from the Kamusi Project (`kamusi.org`), a user-contributed multilingual dictionary. Making use of this resource allowed us to achieve wide lexical coverage quickly, but the heterogeneous nature of user-contributed content also poses some challenges when adapting it for use in an expert system.

Keywords: Swahili, morphological parsing, crowdsourcing

1. Introduction

This paper describes a Swahili morphological analyzer intended for use in statistical machine translation, based on the lexical resources of the user-contributed Kamusi Swahili Dictionary (`kamusi.org`). Adapting these resources into our system gave us significant gains in type and token coverage (+24.61% and +15.70%, respectively), but they also posed challenges, given the heterogeneous nature of user-contributed content.

The use of publicly-contributed data is not an entirely new development in lexicography – the Oxford English Dictionary solicited public contributions as early as 1859 (Winchester, 1999) – but the advent of the internet has made it possible to "crowdsource" (Howe, 2006) a dictionary mostly or entirely from online contributions (Trap-Jensen, 2013). Although recent online crowdsourcing efforts by major English dictionaries like the OED (`oed.com`) and the Collins dictionary (`collinsdictionary.com`) have attracted controversy, most of the world's languages have few if any publicly-available lexical resources in the first place. In such cases, crowdsourcing provides a golden opportunity for wide-coverage but low-cost online resources, whether the languages be major world languages with limited online resources like Swahili, non-standard dialects with limited textual evidence like Philippine English (`www.languagecommunities.com/pinoyenglish.html`), or endangered languages like Inuktitut (`livingdictionary.org`) or Canadian First Nations languages (`firstvoices.org`).

The emergence of these new resources provides a potentially great benefit for those working on the computational processing of these languages (Cristea et al., 2008; Kosem et al., 2013). One question that arises with the respect to community lexicography is whether the resulting dictionaries are of sufficient quality to allow subsequent application development. Regarding the Kamusi dictionary, this project suggests that the answer is certainly "yes", although to some extent the structure of the dictionary and the varying specificity of user-supplied metadata posed some challenges to its adaptation.

The purpose of this article is not to criticize the structure or quality of the Kamusi dictionary, but to illustrate that there is a necessary tension between the following two questions:

1. What sorts of data do developers need when constructing language-related applications and expert systems?
2. What sorts of data are reasonable to ask of community contributors?

The answers to these questions will vary depending on the needs of the application and the nature of the community, but it is almost inevitable that there be some disconnect between what developers would desire and what the "crowd" can deliver. This is particularly true for resources, like the Kamusi dictionary, that are created in advance of the potential projects that their data may end up supporting. In Section 4 we examine some examples in which data crucial to our application (morphological stemming) are not always reasonable to request of community participants working independently.

2. Swahili overview

Swahili is the most widely spoken of the Bantu languages, a branch of the Niger-Kordofanian language family, and is used as a lingua franca across East Africa (Wald, 1990).

The Bantu languages, including Swahili, are well-known for their intricate system of noun classes (Denny and Creider, 1976; Zawawi, 1979; Contini-Morava, 1994). Each noun in Swahili is classified into one of about fourteen "classes" or "genders" . For example, most nouns that describe humans fall into class 1 in the singular, where they are prefixed with *m-* or *mw-*, and fall into class 2 in the plural, where they are prefixed with *wa-*. Most odd-numbered classes have an even numbered class corresponding to its plural; for example, class 3 nouns have class 4 plurals, while class 10 plurals correspond to class 9 or 11 nouns, and some class 11 nouns have class 6 plurals.

Which class a particular noun will belong to is not entirely predictable based on its semantics alone, although there are strong tendencies – humans tend to fall into class 1, plants and extended objects in class 3, abstract ideas in class 11/14, etc. Historically, 22 classes can be reconstructed, but no modern language retains all 22. Table 1 is based on (but,

Singular class	Plural class	Semantic extension
1 (*m-*)	2 (*wa-*)	people
3 (*m-*)	4 (*mi-*)	plants; extended objects; natural phenomena
5 (*0-*, *ji-*)	6 (*ma-*)	fruits; round, hollow, or large things
7 (*ki-*)	8 (*vi-*)	small things
9 (*0-*, *n-*)	10 (*0*, *n-*)	miscellany; loans
11/14 (*u-*)	10 (*0*, *n-*)	masses; abstractions
15 (*ku-*)		gerunds
16, 17, 18 (*-ni*)		locatives

Table 1: Swahili noun classes

of necessity, simplified from) the semantic networks given in (Contini-Morava, 1994).

Verbs, adjectives, demonstratives, possessive markers, relativizers, and other elements all take class affixes as well, to indicate with which nominal arguments they are associated. Identifying these prefixes is therefore essential to identifying correct lexical entries and resolving sentential dependencies, and is the most important task a morphological parser needs to accomplish.

Many of these agreement affixes are homophonous. A *wa-* prefix on a verb, for example, can indicate a 3rd person plural subject, a 2nd singular, 3rd singular, class 3 or class 11 subject in the present tense, or a 2nd plural or 3rd plural object. In concert with other homophonous agreement prefixes, such prefixes can lead to an explosion of ambiguity. Keeping such ambiguity to a minimum was a primary design goal of our system.

3. Extending XFST and LEXC

Our morphological parser serves to break a complex word, such as *ninakupenda* ("I love you") into its component meaningful parts, like [SUBJ.1SG][PRES][OBJ.2SG]pend[VERB]. Given the morphological complexity of Swahili verbs, in particular, this is a necessary step before further processing, as a single verb root may have thousands of possible derivational and inflectional forms.

We implemented our parser using the Xerox finite state toolkit (Beesley and Karttunen, 2003)[1], writing our language description in a higher-level extension of LEXC and compiling this into actual commands in the XFST and LEXC languages.

XFST and LEXC allow the construction of a finite state transducer between, on the upper side, morphological features like [PRES], [SUBJ.1SG], or [VERB], and on the lower side, corresponding segmental material like *-na-*, *ni-*, and *-a*. Morphological co-occurrence is constrained mostly in the LEXC source[2], while phonological rules are specified in the XFST source.

Describing Swahili also requires a number of morphophonological rules: phonological transformations that apply only to particular subsets of stems given the presence or absence of certain morphological features. Class prefixes like *m(w)-* or *ku-*, for example, can behave differently depending on the morphosyntactic class of the following stem. Some sequences, like the negative *ha-* followed by the 1st person singular *ni-*, can coalesce into portmanteau morphemes, but only in particular tenses.

Implementing such morphophonological rules in XFST/LEXC is not entirely straightforward, since they involve lower-side transformations of phonological segments based on the presence or absence of upper-side features like [VERB] or [PAST]. There are several ways of achieving this, but each has some drawbacks.

For example, one can introduce unpronounced "segments" on the lower side that correspond to upper-side features (say, ^PAST^) and then make reference to them in lower-side transformations. One could also use XFST compositions to partition the FST into two parts, apply the lower-side transformation only to one of those parts, and then merge the parts back into one FST. For example, to apply a rule to only [PAST] forms, something like the following is necessary:

```
define Pasts [ $"[PAST]" .o. Lexicon ];
define Nonpasts [ ~$["PAST"] .o. Lexicon];
define Pasts [ Pasts .o. <rule> ];
define Lexicon [ Pasts | Nonpasts ];
```

In this code, one takes the existing lexicon and, using upper-side FST compositions, partitions it between forms that contain [PAST] on their upper sides and forms that do not. One applies the phonological rule only to the former set, using a lower-side composition, and then reconstitutes the lexicon by taking the union of the two FSTs. This can be condensed, of course, but at the expense of readability; the condensed version is even less intuitive, and both are difficult to maintain in the face of changes. In particular, the cascade of partitions and merges that perform a required morphophonological rule differ from those that perform a optional rule, and changing from one to the other can be error-prone.

It is reasonable, when only a few morphophonological rules are necessary, to code each such rule by hand, but Swahili requires dozens of such rules, leading to potentially hundreds of lines of XFST compositions and unions. Instead, we opted to define a very conservative extension to LEXC, in which structured comments allow the user to declare morphophonological transformations more straightforwardly. These comments allow the implementer to define morphological classes using a simple Boolean algebra, and define transformations that obligatorily or optionally apply to them, in a single line in the LEXC file. The following line, for example, performs the *ha-ni-* to *si-* coalescence on the lower side only when the upper side contains [SUBJ.1SG] but does not contain [COND]:

```
!? with [SUBJ.1SG] & ~[COND]
   require {haˆniˆ} -> {siˆ}
```

[1] An earlier, proprietary Swahili parser, written using the TWOLC finite state toolkit, is described in Huskiarnen (1992).

[2] Since Bantu languages, including Swahili, have long-distance restrictions on morpheme co-occurrence, we made extensive use of LEXC flag diacritics, similar to the approach used by Pretorius and Bosch (2003) for Zulu, a related Bantu language.

All the structured comments in the LEXC file are collected and compiled to an XFST script that achieves the appropriate transformation.

There are several advantages to this method. For one, the savings in lines-of-code is substantial, replacing many lines of hard-to-read and hard-to-maintain boilerplate code. It allows the implementer to keep morphophonological declarations in the LEXC file near the morphological declarations they reference, rather than in a separate file.

Furthermore, keywords highlight important distinctions like whether a rule is obligatory ("require") versus optional ("allow"), a critical distinction that is difficult to infer from the corresponding series of XFST compositions. This distinction became very important as we integrated lexical items from the Kamusi dictionary into our system and found that many transformations described in Swahili references as obligatory were actually optional. Changing "require" to "allow" was much quicker, and much less error-prone, than manually editing the code sequences that partition and departition the FST.

This extension also introduces keywords for a few other useful operations. An IF operator allows for the conditional evaluation of rules, useful for enabling and disabling parts of the system for testing. An IMPORT operator allows for the importation of lexical resources into the LEXC file.

One drawback of this particular implementation, however, is that it decouples programmer code from error reporting: the XFST interpreter reports errors in the compiled XFST code, not the extended LEXC code that the programmer works with, meaning that it takes an intermediate step, in the case of error, to determine where the programmer error occurred. A more mature system would perform its own error checking before compilation.

4. Challenges of using user-contributed data

While the Kamusi dictionary made it possible for us to achieve extensive lexical coverage, its user-contributed nature, as well as some aspects of its construction that follow from that, provided the project with certain challenges. Each of the following issues illustrates a tension in crowdsourced lexicography: that information that may be useful or crucial to application developers may be unreasonable or problematic to demand of contributors.

4.1. Senses as entries

Swahili and English word senses do not, of course, correspond perfectly, but lie in a many-to-many correspondence with each other (Table 2). This necessitates the question of which word, Swahili or English, is to be considered the basic entry.

By design, however, the Kamusi dictionary does not privilege any particular language's words as "basic"[3]; nor does it require contributors to agree on which parts of a mapping constitute an "entry". Rather, the dictionary represents a mapping like the above as a series of one-to-one correspondences (Table 3).

A given Swahili word may be spread across many sense-entries, each added by a potentially different contributor.

[3]The Kamusi dictionary is intended to allow correspondences between many languages, not just between Swahili and English.

mtawala	↔	ruler
	↔	administrator
mkabidhi	↔	trustee
	↔	miser
mchopozi	↔	thief
mwizi	↔	

Table 2: Swahili and English senses in a many-to-many correspondence.

mtawala	↔	ruler
mtawala	↔	administrator
mkabidhi	↔	administrator
mkabidhi	↔	trustee
mkabidhi	↔	miser
mchopozi	↔	miser
mchopozi	↔	thief
mwizi	↔	thief

Table 3: Swahili and English senses in a one-to-one correspondence.

This makes it difficult, however, to determine which sets of entries should be considered units by our own system.

Consider *uzazi*, which has twelve sense-entries (Table 4). It is clear that *uzazi* should not be treated as a genuine twelve-way ambiguity, but nor does it represent one unified concept. Rather, there are at least two distinct (albeit related) senses: child-bearing on one hand, and a more abstract concept of familial descent.

In practice, however, we opted to treat all such entries as single entries so long as they had compatible parts of speech. It is likely in some cases that we incorrectly combined genuine homophonies, but this does not affect the primary goal of a morphological stemmer. When these entries differ in their metadata declarations, a question arises that does affect morphological parsing; we will detail this issue in Sections 4.3 and 4.4.

4.2. Lack of morphological breakdowns

The Kamusi contributors do, in fact, perform some stemming, since (in the case of verbs) they are entering the verb

childbirth	delivery
lineage	confinement
descent	fertility
kinship (degree of)	parentage
procreation	propagation
relationship (degree of)	reproduction

Table 4: Twelve senses of "uzazi".

stem, not the fully inflected verb. Beyond that, however, stems are treated as atomic units.

It is unreasonable, in a crowd-sourced project, to expect all (or even most) contributors to be able to perform correct morphological parsing; this is especially true when the morphology is derivational (or even purely etymological). For example, the *-esh* in the stem *onyesha* is a causative suffix, but contributors may not realize this. The root of *-jifunza* ("learn") is in fact *-funza* – the *ji-* is a reflexive prefix – but the root on its own is not common. The Kamusi dictionary therefore does not request that users stem the word beyond removing inflectional morphology.

Entries also do not indicate derivational correspondences between senses – that is, although *-igiza* ("perform") is the causative of *-iga* ("imitate"), there is nothing in the entry of either to indicate that they are related.[4] Such information would, of course, be of great value to a morphological parser.

A morphological parse can often be inferred from the part of speech – an initial *ny-* on a class 9 noun is probably a class 9 prefix, and an *-ish/-esh* suffix in a verbal stem is very likely to be the causative suffix – but the root is not always straightforwardly predictable. Some roots coincidentally begin or end in segments that can be mistaken for affixes, and others take epenthetic segments before particular suffixes. For example, some roots take an epenthetic *-l-* before the causative suffix. If one were to know the corresponding non-causative form, it would become obvious whether the *l* is part of the root or not.

Fortunately, such roots are not especially frequent, and many also occur in our hand-compiled dictionary, so for cases like the epenthetic *-l-* we opted to stem the Kamusi words greedily, always assuming that segments that can be part of an affix are indeed so.

4.3. Specificity of part-of-speech tagging

To stem a Swahili word correctly, it is necessary to know its part of speech. Many noun stems begin with *u*, for example, but one needs to know the class of the noun to know whether this is a class prefix or part of the stem itself. The task of part-of-speech tagging in Swahili is one that can be performed with varying degrees of specificity; some users use broad categories like "noun", others specify the specific class or even subclass.

It is therefore common that a Swahili word, like *uzazi* above, has some sense-entries that specify the noun class and others that merely mark it as a noun. In the case of *uzazi*, three of the senses were tagged as "noun 14", while nine were tagged as "noun". Since *u-* does not uniquely identify the noun class of *uzazi*, there is a potential dilemma here: do we treat this as one "noun 14" entry (with twelve senses), or as one "noun 14" entry (with three senses) and one unclassed "noun" entry (with the remaining nine senses)?[5]

In practice, we treated them as one entry when the following criteria were true:

1. That the proposed parts of speech were compatible. (That is, they form a subset relationship like "noun" and "noun 14", rather than incompatible declarations like "noun 14" and "noun 5".)
2. That the more specific part of speech was morphologically compatible with the form. (That is, that the form of *uzazi* was, in fact, compatible with it being a "noun 14".)

Some users specified class further: not just "noun class 9/10" but "noun class 9/10 animate". Most nouns describing humans are in class 1/2, but some – especially words for occupations – are in other classes. These nouns are associated with their own class agreement prefixes in some paradigms but use class 1/2 agreement prefixes in other. This is useful information for parsing, but many users do not specify it.[6] There are therefore three possibilities for the tagging of a "9/10 animate" noun: "noun", "noun 9/10", and "noun 9/10 animate".

As we did for entries like *uzazi*, we treated such sets as the most specific category, so long as it was morphologically consistent and there were not conflicting declarations. This has the potentiality of introducing another kind of miscategorization; if a form is legitimately ambiguous between an animate (that is, taking class 1/2 agreement morphology for some paradigms) and an inanimate sense, the stemmer would treat it as unambiguously animate.

4.4. Criteria for part-of-speech tagging

Contributors also, at times, use somewhat different notions of part-of-speech. For example, in one sense Swahili "has no prepositions as such" (Erickson and Gustafsson, 1989), and expresses prepositional meanings by means of positional nouns, by an applicative suffix on verbs, and by other means. That is, in many cases where English would use a preposition like "by", Swahili would use a noun meaning "side".

As a result, contributors have a decision to make when assigning words like *kando* ("side, by"), *ndani* ("inside, within"), or *nje* ("outside, external, out of") to parts of speech: do they go by the formal morphological class of the word (in these cases, class 9/10 nouns) or the function of the word in the sentence? A similar issue arises for adjectives and adverbs: many words that function as adjectives and adverbs are not adjectives or adverbs in the formal morphological sense. Since not all contributors will use the same decision criteria, such words may have entries with up to four different parts of speech.

The reason this is problematic for us is that it causes our system to declare a parse ambiguity where none actually exists – the [NOUN], [PREP], [ADJ] and [ADV] parses of a "prepositional" noun are not genuinely different Swahili words, but the result of contributors having to decide between formal and functional descriptions of word usage. We opted to use only the entries labeled as nouns, and for prepositions, adjectives, and adverbs relied instead on hand-chosen lists.

[4] Singular-plural pairs are the only exception; they are treated as one entry.

[5] That is, do parses of sentences containing *uzazi* treat this form as having one potential parse (*u-zazi*) or two (*u-zazi* and *uzazi*)?

[6] This is not surprising, since this information is not apparent from the word form itself, but only by considering the agreement possibilities of its use in a sentence.

Word Class	Custom dict.	Kamusi dict.
Nouns 1/2	48	1177
Nouns 3/4	46	951
Nouns 5/6	41	1762
Nouns 7/8	53	1096
Nouns 9/10	89	3158
Nouns 11/6, 11/10, & 14	36	748
Total classified nouns	**313**	**8892**
Unclassified nouns	–	1860
Total nouns	**313**	**10752**
Verbs of Bantu origin	270	2442
Verbs of Arabic origin	45	542
Total verbs	**315**	**2984**
Prepositions	18	–
Adjectives	72	–
Interrogatives	8	–
Interjections	16	–
Conjunctions	18	–
Punctuation	35	–
Total other	**164**	–
Total words	**792**	**13736**

Table 5: Word entries by class

5. Lexicon

Our morphological parser used lexical data from three sources:

- A small, custom dictionary of Swahili, with roots listed in (and deduced from) textbooks and grammars like Wilson (1983) and Adam (1993).
- A subset of the Kamusi Swahili Dictionary, consisting of nouns with declared classes, unclassed nouns, and verbs. Since classification beyond these classes presents problems of analysis (cf. Section 4.4), we did not include other part-of-speech declarations.
- An English word list (Loge and Beresford, 1999), to catch English loanwords, which are relatively common in online forum speech.

These entry counts are for word forms, not Kamusi dictionary sense entries, so a word like *uzazi* counts as one entry rather than twelve.

6. Tagging "provenance"

Our Swahili morphological parser is intended to provide parsed input to several Swahili-language projects with varying aims, and therefore uses a variety of tags beyond purely morphological features.[7]

Among these are tags intended to signal to downstream modules our confidence in a particular parse. Since our lexical resources are somewhat heterogeneous, we label parses according to the "provenance" – that is, where the stem comes from and how likely we believe that we have inferred the genuine Swahili root. These tags are intended to help later modules in the tool chain choose between ambiguous parses.

- No provenance tag indicates that the root had been collected by hand, usually from textbooks or grammatical resources, and that we had knowledge of, or sufficient information to deduce, the actual root.
- [GUESS1] indicates a root inferred from a Kamusi dictionary stem, but only when its prefixes and suffixes are consistent with its declared part of speech. That is, the stem *uwazi* ("open space, openness") is tagged as [GUESS1] because its class prefix (*u-*) is compatible with its declared part of speech ("noun 14").
- [GUESS2] indicates a unclassified noun from the Kamusi dictionary – one labeled simply as "noun" rather than, say, "noun 1/2". We did not attempt to stem such nouns or predict their class.
- [GUESS3] indicates a possible English loanword.
- [GUESS4] indicates a complete guess. For example, the hypothetical token *kufiva* would have a guessed root *fiv*, which is a possible Swahili root despite it not being listed in any of our lexical resources. [GUESS4] also contains guessed proper names.

In the ambiguity figures that follow in Section 7, we do not count parses of more uncertain provenance as ambiguous parses. (In other words, if a word has two [GUESS1] parses, a [GUESS2] parse, and three [GUESS4] parses, we list the word as having two parses.)

7. Coverage and Ambiguity

Excluding [GUESS4] parses[8], our system exhibits 90.23% token coverage and 66.4% type coverage of a 340k-token Swahili corpus, sourced from the Swahili-language edition of Global Voices (`sw.globalvoicesonline.org`). Coverage percentages are given in Figure 1.

It is worth noting that the inclusion of GUESS level 2, consisting of unclassed nouns for which no more specific class information was available, contributed very little to overall coverage. That is to say, the inclusion of the higher-quality data would have been sufficient, while the inclusion of data of more dubious provenance contributed little.

Each level of guessing contributes greater coverage, but also introduces a greater degree of ambiguity. The more certain we can be that a root is genuine, the better we can reduce illegitimate ambiguities; at guess level 4, where we have little confidence that a parse's root is genuine, it is more likely that any given parse is illegitimate. Figure 2 illustrates the additional degrees of ambiguity contributed by

[7] For example, to allow the inference of stylistic properties of a text, the parser also tags words according to their etymological origin ([ARABIC], [ENG], or a [LOAN] from another language), and whether a particular construction is considered [CASUAL], or tends to occur in [HEADLINE]s.

[8] Including [GUESS4] parses, the system covers 99.78% of tokens and 98.79% of types. The remaining unparsed types are largely usernames or words in other languages and scripts.

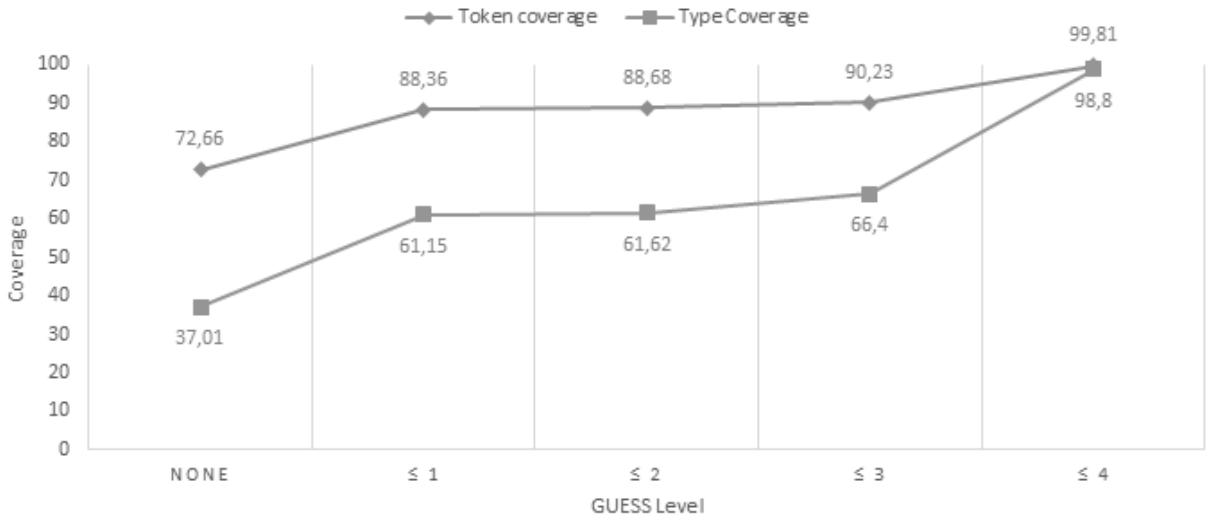

Figure 1: Token and type coverage

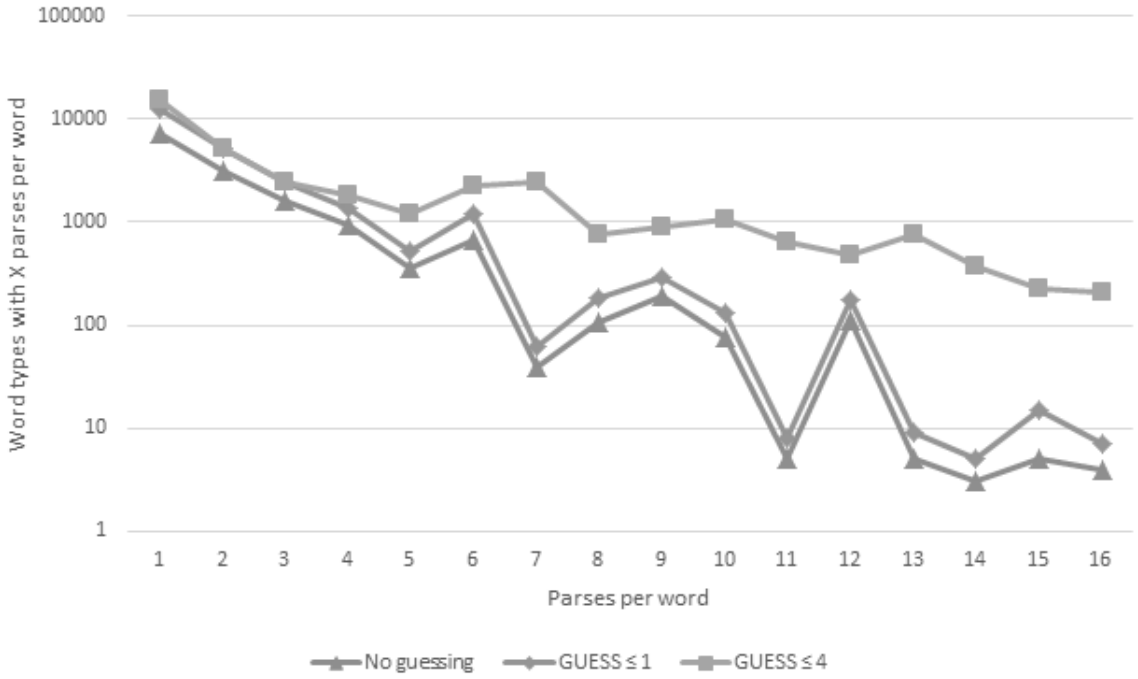

Figure 2: Parse ambiguity

each level of guessing.[9] The X axis represents the number of parses generated by a form, from unambiguous (1 parse) to more ambiguous. The Y axis represents the number of word forms with the given number of parses; for example, at GUESS level 0, there are 7,215 words that are unambiguous, 3,160 words that are two-ways ambiguous, 1,624 words that are three-ways ambiguous, etc.

8. Summary

The Swahili lexical data from the online Kamusi dictionary proved invaluable in extending our system's lexical coverage, but required special care when adapting it as a LEXC resource to minimize ambiguities and illegitimate parses. We sought to contain these in several ways: by excluding lexical forms that appeared to contradict the declared word class, by partitioning the lexical data according to how much word class information was supplied, by supplementing the lexicon with hand-chosen forms, and by tagging our confidence level so that client applications can distinguish the quality of data.

Our experience with the Kamusi dictionary shows that user-contributed data sources are indeed of value when constructing expert systems, and that the advantages of both hand-crafted and crowdsourced resources can be preserved, and the downsides mitigated, when they are used in combination.

[9] GUESS levels 2 and 3 are not included in this table; as they consist of single, unparsed words, they do not contribute to overall ambiguity except when they coincidentally correspond to a GUESS level 0 or 1 word.

9. Acknowledgments

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533.

10. References

Adam, H. (1993). *Kiswahili Intermediate Course*. OMIMEE Intercultural Publishers.

Beesley, K. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications.

Contini-Morava, E. (1994). Noun classification in Swahili. *Publications of the Institute for Advanced Technology in the Humanities, University of Virginia. Research Reports, Second Series*.

Cristea, D., Forscu, C., Rschip, M., and Zock, M. (2008). How to evaluate and raise the quality in a collaborative lexicographic approach. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Denny, J. P. and Creider, C. (1976). The semantics of noun classes in Proto-Bantu. *Studies in African linguistics*, 7(1):1–30.

Erickson, H. and Gustafsson, M. (1989). *Kiswahili Grammar Notes*.

Howe, J. (2006). The rise of crowdsourcing. *Wired*, 14(6).

Huskiarnen, A. (1992). A two-level computer formalism for the analysis of Bantu morphology: An application to Swahili. *Nordic Journal of African Studies*, 1(1):88–122.

Kosem, I., Gantar, P., and Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., and Tuulik, M., editors, *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.*, pages 32–48. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Loge, K. and Beresford, J. R. (1999). The English open word list. http://dreamsteep.com/projects/the-english-open-word-list.html.

Pretorius, L. and Bosch, S. E. (2003). Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation*, pages 191–212.

Trap-Jensen, L. (2013). Researching lexicographical practice. In Jackson, H., editor, *Bloomsbury Companion to Lexicography*, pages 35–47. London: Bloomsbury Academic.

Wald, B. (1990). Swahili and the Bantu languages. In Comrie, B., editor, *The Major Languages of South Asia, the Middle East, and Africa*, pages 219–237.

Wilson, P. (1983). *Simplified Swahili*. Longman Group UK.

Winchester, S. (1999). *The Professor and the Madman*. New York: HarperPerennial.

Zawawi, S. (1979). *Loan words and their effect on the classification of Swahili nominals*. Leiden: E.J. Brill.