Expanding the Lexicon for a Resource-Poor Language Using a Morphological Analyzer and a Web Crawler

Michael Gasser

School of Informatics and Computing, Indiana University
Bloomington, Indiana, USA 47405
gasser@cs.indiana.edu

Abstract

Resource-poor languages may suffer from a lack of any of the basic resources that are fundamental to computational linguistics, including an adequate digital lexicon. Given the relatively small corpus of texts that exists for such languages, extending the lexicon presents a challenge. Languages with complex morphology present a special case, however, because individual words in these languages provide a great deal of information about the grammatical properties of the roots that they are based on. Given a morphological analyzer, it is even possible to extract novel roots from words. In this paper, we look at the case of Tigrinya, a Semitic language with limited lexical resources for which a morphological analyzer is available. It is shown that this analyzer applied to the list of more than 200,000 Tigrinya words that is extracted by a web crawler can extend the lexicon in two ways, by adding new roots and by inferring some of the derivational constraints that apply to known roots.

1. Resource-poor languages and morphological analysis

Resource-poor languages may suffer from a lack of any of the basic resources that are fundamental to computational linguistics, including an adequate digital lexicon. Given the relatively small corpus of texts that exists for such a language, how is it possible to extend the lexicon by adding new lexemes and incorporating information about valency and derivational possibilities into existing lexical entries?

Morphologically complex languages present special problems in this regard because lexemes are not immediately apparent from the wordforms that appear in texts. A morphological analyzer is a fundamental tool for such languages. Machine learning techniques are not yet adequate for the automatic acquisition of morpholological analyzers for very complex languages, especially given data sparsity, but relatively complete grammatical descriptions exist for many languages, and advances in finite state morphology have facilitated the task of creating morphological analyzers for many languages (see Beesley & Karttunen, 2003 for an overview).

A modern morphological analyzer usually incorporates three components into a single finite state transducer: a lexicon of roots or stems; morphotactics, that is, a specification of the order and canonical form for the morphemes in a word; and a set of alternation rules specifying how phonemes, graphemes, and morphemes change in particular contexts. For example, an analysis of the English noun *bosses'* would be based on the presence of the stem *boss* in the lexicon, the morphotactic constraints specifying the order STEM+PLURAL+POSSESSIVE for the morphemes in an English noun, and the phonological/orthographic alternation rule specifying that the plural morpheme is realized as *-es* following a stem ending in *s*.

Obviously a *lexical* morphological analyzer such as this fails when the input word is based on a lexeme that is missing in the lexicon, a situation that is common for resource-poor languages with limited lexical resources. However, if the language has constraints on the shape of roots or stems, it is possible to construct a "guesser" analyzer (Beesley and Karttunen, 2003) which functions like an ordinary lexical analyzer except that it has no lexicon. A guesser analyzer hypothesizes the root/stem for a given input word as a part of analysis.

Given a morphological analyzer, we are in a position to extract a range of other information for both familiar and unfamiliar lexemes, even in the absence of a syntactic parser. In morphologically complex languages, much of the work of syntax is done, or redundantly coded, within words. Combination of roots with various derivational morphemes may correspond to entire phrases in morphologically simpler languages; however, the details of what is possible may be specific to particular lexemes and need to be learned. In addition to the subject-verb agreement that is familiar from Indo-European languages, many languages require verbs to agree with their direct or indirect objects under certain circumstances. Both sorts of agreement can provide information about the syntactic properties of verb lexemes. In the next section, we look at an example of a morphologically complex, resource-poor language and consider how a morphological analyzer could be used to enhance the available verb lexicon.

2. Tigrinya verb morphology

Tigrinya is a language in the Ethio-Semitic family spoken by 5-6 million people in the Tigray region of northern Ethiopia and in central Eritrea. Tigrinya is written in the Ge'ez script common to other languages in the family. There has been almost no computational work on the language, and digital resources are very limited. Although several excellent Tigrinya dictionaries exist, none has been digitized. The only readily available digital lexicon is the ongoing online Tigrinya dictionary project of Efrem Zacarias Zacarias (2009). From the version of that dictionary of February 2008, it was possible to extract only 598 verb roots from among the thousands that are part of the language.

A Tigrinya verb (Leslau, 1941 is a standard reference for Tigrinya grammar) consists of a stem and up to four prefixes and four suffixes: *kemzey_erbIHom*[1] (*kem-zI-ay-yI-**arbIH**-om*) 'that he doesn't benefit (lit. cause to profit) them'. Most of the complexity resides in the stem, which is made up of a root, the only strictly lexical component of the verb, and a "template" representing a combination of tense/aspect/mood (TAM) and one of eight derivational categories.

As in other Semitic languages, the root, consisting of a sequence of consonants, combines with the template, consisting of positions for the root consonants and specific vowels, through a process of "interdigitation" to yield a stem. The template is multimorphemic; it combines one of the four possible TAM categories with one of the eight possible derivational categories. In the word above, the root *rbH* 'profit' combines with the template $aC_1C_2IC_3$, meaning roughly 'present tense, causative', to yield the stem *arbIH* 'cause to profit'.

This root combines with the prefixes *yI* '3rd person masculine singular subject', *ay* 'negative', *zI* 'relativized', and *kem* 'that', and the suffix *om* '3rd person masculine plural object' to yield the word. Note that several phonological changes occur at the boundaries of the morphemes. For example, the sequence *Ia* becomes *e*.

2.1. The HornMorpho Tigrinya analyzer

The root-template interdigitation in the stems of verbs in Tigrinya and other Semitic languages presents challenges to finite state morphology because of its non-concatenative nature (Beesley and Karttunen, 2003). This can be seen as a special case of morphological dependencies spanning intermediate segments. One fruitful approach to the problem of long-distance dependencies in morphology makes use of transducers weighted with feature structure descriptions (Amtrup, 2003). Gasser (2009) shows how this technique can be applied to the analysis and generation of Tigrinya verbs. The current version of the morphological analyzer described in that paper is available as part of the HornMorpho program at `http://www.cs.indiana.edu/~gasser/Research/software.html`.

The prefixes and suffixes within a Tigrinya verb represent subject and object agreement, negation, and relativization, as well as a range of possible prepositions and/or conjunctions. Given the 32 possible stem templates and the possible combinations of affixes, a single Tigrinya verb root can appear in hundreds of thousands of distinct wordforms.

A morphological analyzer should be able to extract the stem from among the affixes and the root and other stem morphemes from within the stem. HornMorpho handles all of the morphological combinations and most of the very complex phonological/orthographic alternations that characterize the Tigrinya verb system. Because of its limited lexical resources, however, the root lexicon of the analyzer is far from adequate. For this reason, HornMorpho has an additional "guesser" analyzer which operates without a lexicon, guessing the roots of verbs for which the lexical analyzer fails.

The HornMorpho Tigrinya analyzer takes as input a Tigrinya word in Ge'ez characters. Each of the Ge'ez characters represents either a consonant vowel sequence or a bare consonant, and romanization is a trivial process of replacing each Ge'ez character with one or two roman characters. However, the Ge'ez writing system does not indicate gemination (consonant lengthening), which, as we will see below, plays a role in the morphology of the language. This introduces some ambiguity, especially when the root of the word is not in the lexicon. Although Ge'ez also fails to indicate the vowel *I*, this presents no particular problem for morphological analysis.

The program first applies its lexical analyzer to the word. If this succeeds, it returns all possible analyses. For example, given the word *zIteKeflel_a* (orthographic *zteKeflela*) 'which was paid for her', the program returns

```
("kfl",
  {tam: prf, der: ps,
   +rel,
   sb: {+3p, -fem, -plr},
   ob: {+3p, +fem, -plr, +prep}})
```

That is, the verb's root is *kfl*, its TAM is perfective, its derivational category is passive, it is relativized, its subject is third person masculine singular, and its indirect (prepositional) object is third person feminine singular.

If the lexical analyzer fails, the program applies the guesser analyzer to the word. For example, for the word above *kemzey_erbIHom* (orthographic *kemzeyerbHom*) , the program returns

```
("rbH",
 {tam: impf, der: cs, conj: kem,
  +rel, +neg,
  sb: {+3p, -fem, -plr},
  ob: {+3p, -fem, +plr}})
```

2.2. Derivational categories and TAM

For all practical purposes, any root can occur with the full range of combinations of prepositions, conjunctions, negation prefix and suffix, relativization suffix, and tense-aspect-mood possibilities. With respect to the derivational categories, however, there may be strict root-specific constraints on what is possible, and knowing which of the categories can occur with which roots greatly simplifies morphological analysis and generation.

Tigrinya verbs fall into eight derivational categories. For the purposes of this paper, however, we will consider only the five most common, which I will refer to as **simplex**, **passive/reflexive**, **transitive/causative**, **reciprocal1**, and **reciprocal2**, though these names do not always accurately reflect the actual semantics. With respect to their derivational possibilities, Tigrinya verb roots fall into two basic classes:

A Roots whose "basic" form is the simplex form. When this form is transitive, the passive/reflexive form of the

[1] In this paper, all Tigrinya words are romanized, following the SERA romanization scheme (Firdyiwek and Yaqob, 1997). A segment followed by a *W*, for example, *kW*, represents a singular labialized consonsant. The underscore character is used to represent gemination (not a SERA convention).

root represents the genuine passive or reflexive; when the simplex form is intransitive, the passive/reflexive form, which occurs rarely if at all, represents an "impersonal passive." Most roots fall into this category. An example is the root *flT* 'know': simplex *feleTe* 'he knew', passive/reflexive *tefelTe* 'he was known', transitive/causative *afleTe* 'he caused to know, informed'.

B Roots which never occur in the simplex form. For most such roots the "basic" form is the passive/reflexive. This form is not a genuine passive or reflexive and may even be transitive. The basic form is usually intransitive, however, and the corresponding transitive form is represented by the transitive/causative form of the root. An example is the root *qm_T* 'sit': passive/reflexive *teQem_eTe* 'he sat', transitive/causative *aQem_eTe* 'he caused to sit, put'. In some cases, it is the reciprocal1 form that is basic; neither the simplex nor the passive/reflexive occurs. An example is the root *kt'* 'argue': reciprocal1 *teKat'E* 'he argued'.

All Tigrinya verbs must agree with their subjects in person, number, and gender; the language makes ten distinctions for different combinations of values on these dimensions. In addition, verbs with a definite direct object must agree with their object (except in limited circumstances which will not be described here), taking one of a set of ten object suffixes. Subject and object agreement will prove useful when we need to examine the variety of environments a root occurs in.

2.3. Root categories

Semitic verbs are complicated further by the fact that the roots fall into categories that interact differently with the TAM and derivation templates. A full discussion of the root categories is beyond the scope of this paper. What is relevant for our purposes is the potential for confusion among several of the categories. In Tigrinya, as in other Ethio-Semitic languages, but not in other Semitic languages, there is a lexical distinction between roots of the form $C_1C_2C_3$, those of the form $C_1C_2\_C_3$, and those of the form $C_1aC_2C_3$. However, the distinction between these categories is obscured in some of the templates.

Consider an invented word whose orthographic form is romanized as *desene*. In the absence of lexical knowledge, this could represent the third person masculine singular perfective form of the root *dsn* or the root *ds_n*. The orthographic word in the two cases would be distinguished phonetically (*desene* vs. *des_ene*), but the orthography fails to indicate the gemination of the second stem consonant.

The distinction between the categories matters, however, for orthography as well as phonology. The imperfective simplex template is $C_1eC_2(\_I)C_3$ for the $C_1C_2C_3$ root category and $C_1 1C_2\_IC_3$ for the $C_1C_2\_C_3$ category. In the third person masculine singular imperfective our imaginary root would take the form *ydesn* (phonetically *yIdes_In*) for the root *dsn* and *ydsn* (phonetically *yIdIs_In*) for the root *ds_n*.

Because of a range of complicated phonological alternations, other sorts of ambiguities can occur, especially when one of a root's consonants is *y* or when two consecutive consonants are the same. A root *y* may end up realized as a vowel in certain templates. For example, the invented word *temege* could be the third person masculine singular perfective passive form of a verb with the root *mg_y* or the third person masculine singular perfective simplex form of a verb with the root *tmg*.

When the same consonant appears twice in succession in a root, the two segments are merged into a single geminated consonant in some templates. Because the gemination is not indicated in the orthography, one of the root consonants is effectively lost. For example, the invented word *kgeru* could be the third person masculine plural imperfective simplex form of a verb with the root *grr* and the conjunctive prefix *k-* or the second person masculine plural imperative simplex form of a verb with the root *kgr*.

In summary, the Tigrinya verb lexicon could be enhanced in two ways using a morphological analyzer: through the addition of new roots and constraints on the cooccurrence of roots with the core derivational categories. The matter is not so simple, however. There is a great deal of ambiguity in the system, especially given relatively little lexical knowledge. Many words which are not even verbs can be analyzed as verb forms based on possible unfamiliar roots. The upshot is that extracting lexical information using morphological analysis is a noisy process. That is, on the basis of a single instance of a word that is analyzed as a verb with an unfamiliar root, we can not thereby conclude that this is an actual root in the language. Therefore, in order to extract the lexical information that we want, the morphological analyzer must be used in combination with a large amount of data. In the next section, we see how a web crawler provides the data that we need.

3. Using a web crawler to extract lexical information

Biniam Gebremichael (2009) has written a web crawler for extracting Tigrinya texts from the Internet. The output of his program is a list of 227,984 unique wordforms, along with their frequencies. Because of the relative complexity of verbs, most of the word types in the language are verbs, so the crawler output provides a great deal of implicit information about the Tigrinya verb lexicon.

After eliminating words of fewer than four characters (unlikely to be verbs), words of more than 14 characters (mostly likely two words with a missing space), and a small number of words known to be non-verbs, we are left with 206,921 words. The HornMorpho analyzer was applied to all of these words. The lexical analyzer succeeded on 65,732 words, and the guesser analyzer gave at least one analysis for 46,979 of the remaining words.

3.1. Inferring Properties of Known Roots

578 of the 598 roots in the analyzer's dictionary are within the analyses of the words that the lexical analyzer succeeds on. Most of these roots appear in multiple words, so it should be possible to infer some of their morphosyntactic properties. The first goal was to split the roots into categories A and B, based on whether they appear in the simplex form.

For each root, its occurrence in each of the eight derivational categories was counted, and any root for which the simplex category represented less than 10% of the total was counted as a member of category B. 57 of the 578 roots fell into this category. Within this category, the basic form for most roots is the passive/reflexive. Those eight roots for which the reciprocal1 category represented more than 50% of the occurrences were counted in the special category of basic reciprocal verbs, and, surprisingly, a further three roots occurred overwhelmingly in the transitive/causative rather than the passive/reflexive or reciprocal1.

A relatively comprehensive Tigrinya-English dictionary (Kane, 2000) was consulted to evaluate the results. It is conventional in Ethio-Semitic dictionaries to distinguish verbs that fail to occur in the simplex form (what is referred to as category B here). Of the 57 roots that were assigned to category B based on the analyzer output, 42 were indicated as such in the dictionary (86%). The fact that the dictionary cites simplex forms for the remaining 15 roots does not mean that these forms occur with any particular frequency. Thus the precision is at least 86%. Of the 521 roots assigned to category A, the dictionary lists only one as failing to occur in the simplex form. That is, recall is 97.5% (42/43).

Within category A, it would also be useful to distinguish roots that are transitive from those that are intransitive in their simplex form. Two sorts of information are relevant here: the occurrence or non-occurrence of the passive/reflexive form and the rate of occurrence of particular object agreement suffixes on the verbs with these roots. The latter information proved not particularly helpful. The problem is that the object suffixes appear on many intransitive verbs with a dative or experiencer function, for example, *yImeslen_i* 'it appears (to) me'.

The occurrence of the passive/reflexive proved more useful, with significant differences across the category A roots. Somewhat surprisingly, based on the proportion of words in the passive/reflexive form for each of the category A roots, no clear dividing line separating intransitive and transitive verbs emerged. Transitivity, at least as measured in this way, appears to be a more or less continuous phenomenon in Tigrinya.

3.2. Inferring New Roots

The output for the words analyzed by the guesser analyzer was extremely messy. Some words had more than 50 analyses, and many non-verbs were analyzed as verbs. For example, the word *fereQa*, a noun meaning 'half', has the following analysis (along with four others):

```
("frq",
 {tam: prf, der: smp,
  -rel, -neg,
  sb: {+3p, -fem, +plr},
  ob: {+3p, +fem, -plr}})
```

Genuine verbs also had multiple analyses. One source of confusion is the possibility in some cases of treating the consonant in a prefix as part of the root. For example, the verb *zIgWet_u* (orthographically *zgWetu*) 'which they pull' from the actual root *gWtt* has the following analysis, along with the correct one and six other incorrect ones:

```
("zgWt",
 {tam: imprv, der: smp,
  -neg,
  sb: {+2p, -fem, +plr}, ob: {}})
```

Here the relativization prefix *z-* is treated as the initial consonant of a hypothesized root in a non-relativized imperative verb.

Although these analyses are "correct" in the sense that they obey the morphotactics and the orthographic and phonological rules of the language, they are also obviously wrong. A total of 25,263 new roots were hypothesized by the guesser analyzer. An informal inspection of the analyses made it clear that the great majority of these roots were not valid.

Roots which the analyzer hypothesized only infrequently are obviously less likely to be valid, so only those appearing in 15 or more analyses were considered further. Many of the roots apparently also consisted of impossible sequences of consonants. In the absence of an account of the constraints on Tigrinya root segments, a trigram model of roots was constructed on the basis of the 598 roots in the HornMorpho lexicon. Based on this model, any root with a negative log probability of greater than 50 was eliminated from the list of candidates. After these two steps, 1529 roots remained.

In order to filter the list of candidate roots further, an attempt was made to discover what range of morphological environments a "good" Tigrinya verb root occurred in. Based on the output of the lexical analyzer, various properties of the known roots were considered. Three sorts of properties proved to be typical. First, roots tend to occur in all four of the TAM categories. Second, roots tend to occur in both relativized and non-relativized forms.[2] Third, roots tend to occur with a wide range of subject and/or object agreement affixes from among the ten that are possible for each.

Based on the typical pattern of occurrence of these features, candidate roots were considered further only

1. if they occurred in at least three of the four TAM categories
2. if their total number of subject and object agreement categories was greater than 4
3. if they occurred in a relativized verb at least 7% of the time.

After this step, 1115 candidate roots remained.

An examination of this set of candidates revealed that it contained a number of pairs of the sort $C_1C_2C_3/C_1C_2\_C_3$, $C_1C_2C_3/C_1aC_2C_3$, or $C_1C_2\_C_3/C_1aC_2C_3$ Further investigation indicated that such pairs were often part of alternate analyses of the same word. For example, the word *bedihom* has these analyses:

```
("bdh",
```

[2] Relative clauses are much more common in Tigrinya than in a language such as English; they often correspond to adjectives in other languages: *zIdereQe 'InCeyti* 'dry wood', literally 'wood which dried'.

		Assignments	Precision
New roots	(after confusable root competition)	417	0.66
	(with confusable roots combined)		0.84
Derivational classes		578	0.86

Table 1: Performance of Morphological Analyzer on Crawler Data

```
 {tam: ger, der: smp,
  -rel, -neg,
  sb: {+3p, -fem, +plr}, ob: {}})

("bd_h", {tam: ger, der: smp,
  -rel, -neg,
  sb: {+3p, -fem, +plr}, ob: {}})
```

In many of these cases, it was apparent that the morphosyntactic criteria used to eliminate roots that deviated from the typical root profile failed to allow a strong preference for either element of these pairs of similar candidate roots. Since the candidates in each pair provided alternate analyses for many of the same words, they could be seen as competing with one another. That is, for each such pair, only one root should survive in the final list of inferred roots.

As a first step towards implementing this competition, a list of competing sets of roots was extracted from the original analyses. A competition set consisted of all of the candidate roots in the list of 1115 that were shared by particular words in the original analysis. For example, since *bdh* and *bd_h* were both considered to be possible roots for the word *bedihom*, they belonged to the same competition set (in fact to several overlapping sets).

A total of 678 competition sets was extracted in this way. The sets consisted of as few as two words and as many as 19. In many cases, they contained the expected grouping of $C_1C_2C_3$, $C_1C_2\_C_3$, and $C_1aC_2C_3$. For example, one set consisted of {*frm*, *fr_m*, *farm*}. Others contained roots that were confusable for other reasons, for example, because a verb prefix was treated as part of a root, a root contained a *y* that was realized as a vowel in the word, or a root contained the same consonant twice in succession. For example, one competition set contained the roots *zmd*, *mmd*, and *myd*, among others.

For each competition set, the root with the highest trigram probability was selected over the other competitors. This step could end up excluding a number of genuine roots, but we are more interested in precision than recall in inferring new roots. After this step, there were 417 roots in the final set of candidates.

To evaluate this list, each of the roots was looked up in a Tigrinya-English dictionary (Kane, 2000). The results are shown in Table 1. A total of 275 of the candidate roots (66%) were deemed by the dictionary to be genuine Tigrinya verb roots. While this is an impressive number of inferred roots, the technique in this form remains largely unusable because of the large number of invalid roots that are included.

3.3. Novel Root Error Analysis

An error analysis revealed that, within the 142 errors, 75 consisted of roots which corresponded to genuine roots with a change of root category but not root consonants. For example, the invalid candidate root *rs_n* corresponded to the genuine root *rsn*, and the invalid candidate root *bags* corresponded to the genuine root *bg_s*. The problem is obviously in the root competition step; the trigram root model fails to prefer genuine roots over similar invalid ones.

The inadequacy of the trigram model may be due to several factors. First, there is relatively little data to work with, only 598 known roots. Except at the beginnings and ends of the roots, backoff must almost always be relied on. Second, a segment-based model cannot make generalizations based on phonetic similarities between segments. For example, there would be no way to conclude from the frequency of a *q* in a particular environment that *k* (which shares velar place of articulation and voicelessness with *q*) should also be relatively acceptable in that environment. Third, no ngram model can make generalizations based on abstract sameness. No Tigrinya verb roots consist of sequences of more than two identical consonants; yet the trigram model assigns a relatively high probability to the candidate root *rrr*.

Even with an adequate model of root phonotactics, it may not be possible to rank the roots within a competition set. That is, it may simply be a lexical accident that *rsn* is an actual root and *rs_n* is not. One could even make an argument that, within some gross constraints about root phonotactics, the most likely element in a competition is the one that *least* resembles existing roots because this would make it maximally distinguishable.

The solution would seem to be to remain neutral about which root form is to be preferred until more information is available. That is, since both *rsn* and *rs_n* are compatible with the existing data from the crawler, we would treat them as a single verb root. Further disambiguating evidence would then lead us to prefer one form over the other, or, in relatively rare cases, to treat them as separate existing roots.

With the 75 errors of this type considered correct, precision rises to 84%.

The remaining errors belong to several categories. A few are variants of very common existing roots; these seem to be the result of common typographical errors. For example, one candidate root is *rk_b*, similar to the genuine root *rkb* 'find', one of the most common roots in the language. Typographical errors in Ge'ez tend to replace one vowel with another, and a simple change of vowels would lead the lexical analyzer to fail and the guesser analyzer to posit a root such as *rk_b*.

Most of the remaining errors consist of candidate roots end-

ing in *y*; many of these prove to result from a single bug in the portion of the morphological analyzer that handles the phonological changes related to root *y*. Another subset of the errors is due to phonological/orthographic variants that are not captured in the current version of HornMorpho. Seven of the errors remain unexplained.

4. Conclusions

We conclude that a morphological analyzer in combination with a list of wordform types output by a language-specific web crawler has the potential to enhance the lexical resources for a morphologically rich but resource-poor language, both increasing the store of known lexemes and adding key morphosyntactic information to the lexemes that are already known. The analysis also has the potential to turn up some unexpected properties of roots, for example, the occurrence of certain Tigrinya roots overwhelmingly in the transitive/causative form and the lack of a clear separation between intransitive and transitive roots in the language. Because of the large number of inherent ambiguities with a language such as Tigrinya, a great deal of filtering is necessary to settle on a set of inferred roots. In this process, the notion of a typical root profile proved useful. The profile was constructed through an informal examination of the morphosyntactic properties of known roots. In future work, we plan to build a root classifier, using the morphosyntactic properties discussed in this paper as well a trigram model based on phonetic features. In addition to continuing work with Tigrinya, these same ideas are being applied to the closely related Semitic language Amharic and the very different indigenous South American language Quechua.

5. References

Jan Amtrup. 2003. Morphology in machine translation systems: Efficient integration of finite state transducers and feature structure descriptions. *Machine Translation*, 18:213–235.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA, USA.

Yitna Firdyiwek and Daniel Yaqob. 1997. The system for Ethiopic representation in ASCII. URL: citeseer.ist.psu.edu/56365.html.

Michael Gasser. 2009. Semitic morphological analysis and generation using finite state transducers with feature structures. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 309–317, Athens, Greece.

Biniam Gebremichael. 2009. Wordlist and spell checking for Amharic and Tigrigna. Available online at http://www.cs.ru.nl/ biniam/geez/crawl.php.

Thomas L. Kane. 2000. *Tigrinya-English Dictionary*. Dunwoody Press, Springfield, VA, USA.

Wolf Leslau. 1941. *Documents Tigrigna: Grammaire et Textes*. Libraire C. Klincksieck, Paris.

Efrem Zacarias. 2009. Memhir.org dictionaries (English-Tigrinya, Hebrew-Tigrinya dictionaries). Available at http://www.memhr.org/dic/.