Part of Speech tagging for Amharic using Conditional Random Fields

Sisay Fissaha Adafre
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
sfissaha@science.uva.nl

Abstract

We applied Conditional Random Fields (CRFs) to the tasks of Amharic word segmentation and POS tagging using a small annotated corpus of 1000 words. Given the size of the data and the large number of unknown words in the test corpus (80%), an accuracy of 84% for Amharic word segmentation and 74% for POS tagging is encouraging, indicating the applicability of CRFs for a morphologically complex language like Amharic.

1 Introduction

Part-of-speech (POS) tagging is often considered as the first phase of a more complex natural language processing application. The task is particularly amenable to automatic processing. Specifically, POS taggers that are trained on pre-annotated corpora achieve human-like performance, which is adequate for most applications. The road to such high performance levels is, however, filled with a hierarchy of sub-problems. Most techniques generally assume the availability of large POS annotated corpora. The development of annotated corpora in turn requires a standard POS tagset. None of these resources are available for Amharic. This is due mainly to the fact that data preparation, i.e., developing a comprehensive POS tagset and annotating a reasonably sized text, is an arduous task. Although the POS tagging task, taken as a whole, seems challenging, a lot can be gained by analyzing it into subproblems and dealing with each one step-by-step, and also bringing in the experience from other languages in solving these problems, since POS taggers have been developed for several languages resulting in a rich body of knowledge.

Several attempts have been made in the past to develop algorithms for analyzing Amharic words. Among these is the stemming algorithm of Nega (1999), which reduces Amharic words into their common stem forms by removing affixes. Nega's work focuses on investigating the effectiveness of the stemming algorithm in information retrieval for Amharic. Abyot (2000) developed a word parser for Amharic verbs that analyses verbs into their constituting morphemes and determines their morphosyntactic categories. Abyot's work only covers verbs and their derivations. Mesfin (2001) developed a Hidden Markov Model (HMM) based part of speech tagger for Amharic. Building on the work of Mesfin, Atelach (2002) developed a stochastic syntactic parser for Amharic. Sisay and Haller (2003a; 2003b) applied finite-state tools, and corpus-based methods for the Amharic morphological analysis. This work provided important insights into the issues surrounding the development of Amharic natural language processing applications, especially, in compiling a preliminary POS tagset for Amharic.

In this paper, our aim is to explore recent developments in the morphological analysis of related languages, such as Arabic and Hebrew, and machine learning approaches, and apply them to the Amharic language. Amharic belongs to the Semitic family of languages, and hence shares a number of common morphological properties with Arabic and Hebrew for which active research is being carried out. Stud-

Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 47–54,
Ann Arbor, June 2005.

ies on these languages propose two alternative POS tagging approaches which differ on the unit of analysis chosen; morpheme-based and word-based (Bar-Haim et al., 2004). The former presupposes a segmentation phase in which words are analysed into constituting morphemes which are then passed to the POS tagging step, whereas the latter applies POS tagging directly on fully-inflected word forms. Due to scarce resources, it is impossible for us to fully carry out these tasks for Amharic. Therefore, the segmentation and POS tagging tasks are carried out independently. Furthermore, POS tagging is applied only on fully-inflected word forms. The motivation for doing the segmentation task comes from the need to provide some measure of the complexity of the task in the context of the Amharic language. As regards implementation, new models have been introduced recently for segmentation and sequence-labeling tasks. One such model is Conditional Random Fields (CRFs) (Lafferty et al., 2001). In this paper, we describe important morphosyntactic characteristics of Amharic, and apply CRFs to Amharic word segmentation and POS tagging.

The paper is organized as follows. Section 2 provides a brief description of Amharic morphology. Section 3 presents some of the work done in the area of Amharic morphological analysis, and examines one POS tagset proposed by previous studies. This tagset has been revised and applied on a sample Amharic newspaper text, which is discussed in Section 4. Section 5 describes the tasks in greater detail. Section 6 provides a brief description of CRFs, the machine learning algorithm that will be applied in this paper. Section 7 describes the experimental setup and Section 8 presents the result of the experiment. Finally, Section 9 makes some concluding remarks.

2 Amharic Morphology

Amharic is one of the most widely spoken languages in Ethiopia. It has its own script that is borrowed from Ge'ez, another Ethiopian Semitic language (Leslau, 1995). The script is believed to have originated from the South Sabean script. It is a syllabary writing system where each character represents an open CV syllable, i.e., a combination of a consonant followed by a vowel (Daniels, 1997).

Amharic has a complex morphology. Word formation involves prefixation, suffixation, infixation, reduplication, and Semitic stem interdigitation, among others. Like other Semitic languages, e.g., Arabic, Amharic verbs and their derivations constitute a significant part of the lexicon. In Semitic languages, words, especially verbs, are best viewed as consisting of discontinuous morphemes that are combined in a non-concatenative manner. Put differently, verbs are commonly analyzed as consisting of root consonants, template patterns, and vowel patterns. With the exception of very few verb forms (such as the imperative), all derived verb forms take affixes in order to appear as independent words.

Most function words in Amharic, such as Conjunction, Preposition, Article, Relative marker, Pronominal affixes, Negation markers, are bound morphemes, which are attached to content words, resulting in complex Amharic words composed of several morphemes. Nouns inflect for the morphosyntactic features number, gender, definiteness, and case. Amharic adjectives share some morphological properties with nouns, such as definiteness, case, and number. As compared to nouns and verbs, there are fewer primary adjectives. Most adjectives are derived from nouns or verbs. Amharic has very few lexical adverbs. Adverbial meaning is usually expressed morphologically on the verb or through prepositional phrases. While prepositions are mostly bound morphemes, postpositions are typically independent words.

The segmentation task (cf. Section 7.1) considers the following bound morphemes as segments: Prepositions, Conjunctions, Relative Makers, Auxiliary verbs, Negation Marker and Coordinate Conjunction. Other bound morphemes such as definite article, agreement features (i.e., number, gender), case markers, etc are not considered as segments and will be treated as part of the word. These are chosen since they are commonly treated as separate units in most syntactic descriptions.

Although the above description of Amharic is far from complete, it highlights some of the major characteristics of Amharic, which it shares with other Semitic languages such as Arabic. It is, therefore, worthwhile to take into consideration the work done for other Semitic languages in proposing a method for Amharic natural language processing.

3 Amharic POS Tagset

Mesfin (2001) compiled a total of 25 POS tags: N, NV, NB, NP, NC, V, AUX, VCO, VP, VC, J, JC, JNU, JPN, JP, PREP, ADV, ADVC, C, REL, ITJ, ORD, CRD, PUNC, and UNC. These tags capture important properties of the language at a higher level of description. For example, the fact that there is no category for Articles indicates that Amharic does not have independent lexical forms for articles. However, a close examination of the description of some of the tags reveals some missclassification that we think will lead to tagging inconsistency. For example, the tag JPN is assigned to nouns with the "ye" prefix morpheme that function as an adjective, e.g. `yetaywan sahn - A Taiwan made plate` (Mesfin, 2001). This example shows that grammatical function takes precedence over morphological form in deciding the POS category of a word. In Amharic, the ye+NOUN construction can also be used to represent other kinds of relation such as Possession relation. On the other hand, the ye+NOUN construction is a simple morphological variant of the NOUN that can easily be recognized. Therefore, treating ye+NOUN construction as a subclass of a major noun class will result in a better tagging consistency than treating it as an adjective. Furthermore, a hierarchical tagset, organized into major classes and subclasses, seems to be a preferred design strategy (Wilson, 1996; Khoja et al., 2001). Although it is possible to guess (from the tagset description) some abstract classes such as, N* (nouns), V* (verbs), J* (adjectives), etc., such a hierarchical relation is not clearly indicated. One advantage of such a hierarchical organization is that it allows one to work at different levels of abstraction.

The POS tags that are used in this paper are obtained by collapsing some of the categories proposed by Mesfin (2001). The POS tags are Noun (N), Verb (V), Auxiliary verbs (AUX), Numerals (NU), Adjective (AJ), Adverb (AV), Adposition (AP), Interjection (I), Residual (R), and Punctuation (PU). The main reason for working with a set of abstract POS tags is resource limitation, i.e., the absence of a large annotated corpus. Since we are working on a small annotated corpus, 25 POS tags make the data sparse and the results unreliable. Therefore, we have found it necessary to revise the tagset.

4 Application of the Revised Tagset

The above abstract POS tags are chosen by taking into account the proposals made in Amharic grammar literature and the guidelines of other languages (Baye, 1986; Wilson, 1996; Khoja et al., 2001). It is, however, necessary to apply the revised tagset to a real Amharic text and see if it leads to any unforeseeable problems. It is also useful to see the distribution of POS tags in a typical Amahric newspaper text. Therefore, we selected 5 Amharic news articles and applied the above tagset.

All the tokens in the corpus are assigned one of the tags in the proposed tagset relatively easily. There do not seem to be any gaps in the tagset. Unlike Mesfin (2001), who assigns collocations a single POS tag, we have assumed that each token should be treated separately. This means that words that are part of a collocation are assigned tags individually. This in turn contributes towards a better tagging consistency by minimizing context dependent decision-making steps.

Table 1 shows the distribution of POS tags in the corpus. Nouns constitute the largest POS category in the corpus based on the above tagging scheme. This seems to be characteristic of other languages too. However, Amharic makes extensive use of noun clauses for representing different kinds of subordinate clauses. Noun clauses are headed by a verbal noun, which is assigned a noun POS tag. This adds to the skewedness of POS tag distributions which in turn biases the POS tagger that relies heavily on morphological features as we will show in Section 7. Interjections, on the other hand, do not occur in the sample corpus, as these words usually do not appear often in newspaper text.

Once the POS tagset has been compiled and tested, the next logical step is to explore automatic methods of analyzing Amharic words, which we explore in the next section.

5 POS Tagging of Amharic

Semitic languages like Arabic, Hebrew and Amharic have a much more complex morphology than English. In these languages, words usually consist of several bound morphemes that would normally have independent lexical entries in languages like English. Furthermore, in Arabic and Hebrew, the

Description	POS tag	Frequency
Noun	N	586
Verb	V	203
Auxiliary	AUX	20
Numeral	NU	65
Adjective	AJ	31
Adverb	AV	8
Adposition	AP	30
Interjection	I	0
Punctuation	PU	36
Residual	R	15

Table 1: Distribution of POS tags

diacritics that represent most vowels and gemination patterns are missing in written texts. Although Amharic does not have a special marker for gemination, the Amharic script fully encodes both the vowels and the consonants, hence it does not suffer from the ambiguity problem that may arise due to the missing vowels.

As mentioned briefly in Section 1, the morphological complexity of these languages opens up different alternative approaches in developing POS taggers for them (Bar-Haim et al., 2004; Diab et al., 2004). Bar-Haim et al. (2004) showed that morpheme-based tagging performs better than word-based tagging; they used Hidden Markov Models (HMMs) for developing the tagger.

On the basis of the idea introduced by Bar-Haim et al. (2004), we formulate the following two related tasks for the analysis of Amharic words: segmentation and POS tagging (sequence labeling). Segmentation refers to the analysis of a word into constituting morphemes. The POS tagging task, on the other hand, deals with the assignment of POS tags to words. The revised POS tags that are introduced in Section 3 will be used for this task. The main reason for choosing words as a unit of analysis and adopting the abstract POS tags is that the limited resource that we have prohibits us from carrying out fine-grained classification experiments. As a result of this, we choose to aim at a less ambitious goal of investigating to what extent the strategies used for unknown word recognitions can help fill the gap left by scarce resources. Therefore, we mainly focus on word-based tagging and explore different kinds of features that contribute to tagging accuracy.

Although the segmentation and POS tagging tasks look different, both can be reduced to sequence labeling tasks. Since the size of the annotated corpora is very small, a method needs to be chosen that allows an optimal utilization of the limited resources that are available for Amharic. In this respect, CRFs are more appropriate than HMMs since they allow us to integrate information from different sources (Lafferty et al., 2001). In the next section, we provide a brief description of CRFs.

6 Conditional Random Fields

Conditional Random Fields are conditional probability distributions that take the form of exponential models. A special case of CRFs, linear chain CRF, which takes the following form, has been widely used for sequence labeling tasks.

$$P\left(y \mid x\right) = \frac{1}{Z\left(x\right)} exp\left(\sum_{t=1}\sum_{k} \lambda_k f_k\left(t, y_{t-1}, y_t, x\right)\right),$$

where $Z\left(x\right)$ is the normalization factor, $X = \{x_1, \ldots, x_n\}$ is the observation sequence, $Y = \{y_1, \ldots, y_T\}$ is the label sequences, f_k and λ_k are the feature functions and their corresponding weights respectively (Lafferty et al., 2001).

An important property of these models is that probabilities are computed based on a set of feature functions, i.e. f_k, (usually binary valued), which are defined on both the observation X and label sequences Y. These feature functions describe different aspect of the data and may overlap, providing a flexible way of describing the task. CRFs have been shown to perform well in a number of natural language processing applications, such as POS tagging (Lafferty et al., 2001), shallow parsing or NP chunking (Sha and Pereira, 2003), and named entity recognition (McCallum and Li, 2003).

In POS tagging, context information such as surrounding words and their morphological features, i.e., suffixes and prefixes, significantly improves performance. CRFs allow us to integrate large set of such features easily. Therefore, it would be interesting to see to what extent the morphological features help in predicting Amharic POS tags. We used the minorThird implementation of CRF (Cohen, 2004).

7 Experiments

There are limited resources for the Amharic language, which can be used for developing POS tagger. One resource that may be relevant for the current task is a dictionary consisting of some 15,000 entries (Amsalu, 1987). Each entry is assigned one of the five POS tags; Noun, Verb, Adjectives, Adverb, and Adposition. Due to the morphological complexity of the language, a fully inflected dictionary consisting only of 15,000 entries is bound to have limited coverage. Furthermore, the dictionary contains entries for phrases, which do not fall into any of the POS categories. Therefore the actual number of useful entries is a lot less than 15,000.

The data for the experiment that will be described below consists of 5 annotated news articles (1000 words). The Amharic text has been transliterated using the SERA transliteration scheme, which encodes Amharic scripts using Latin alphabets (Daniel, 1996). This data is very small compared to the data used in other segmentation and POS tagging experiments. However, it is worthwhile to investigate how such a limited resource can meaningfully be used for tackling the aforementioned tasks.

7.1 Segmentation

The training data for segmentation task consists of 5 news articles in which the words are annotated with segment boundaries as shown in the following example.

```
...<seg>Ind</seg><seg>
astawequt</seg>#
<seg>le</seg><seg>arso
</seg>#<seg> aderu
</seg># <seg>be</seg>
<seg>temeTaTaN</seg>...
```

In this example, the morphemes are enclosed in `<seg>` and `</seg>` XML tags. Word-boundaries are indicated using the special symbol #. The reduction of the segmentation task to a sequence labeling task is achieved by converting the XML-annotated text into a sequence of character-tag pairs. Each character constitutes a training (test) instance. The following five tags are used for tagging the characters; B(egin), C(ontinue), E(nd), U(nique) and N(egative). Each character in the segment is assigned one of these tags depending on where it appears in the segment; at the beginning (B), at the end (E), inside (C), or alone (U). While the tags BCE are used to capture multi-character morphemes, the U tag is used to represent single-character morphemes. The negative tag (N) is assigned to the special symbol # used to indicate the word boundary. Though experiments have been carried out with less elaborate tagging schemes such as BIO (Begin-Inside-Outside), no significant performance improvement has been observed. Therefore, results are reported only for the BCEUN tagging scheme.

The set of features that are used for training are composed of character features, morphological features, dictionary features, the previous tag, and character bi-grams. We used a window of eleven characters centered at the current character. The character features consist of the current character, the five characters to the left and to the right of the current characters. Morphological features are generated by first merging the set of characters that appear between the word boundaries (both left and right) and the current character. Then a binary feature will be generated in which its value depends on whether the resulting segment appears in a precompiled list of valid prefix and suffix morphemes or not. The same segment is also used to generate another dictionary-based feature, i.e., it is checked whether it exists in the dictionary. Character bi-grams that appear to the left and the right of the current character are also used as features. Finally, the previous tag is also used as a feature.

7.2 POS Tagging

The experimental setup for POS tagging is similar to that of the segmentation task. However, in our current experiments, words, instead of characters, are annotated with their POS tags and hence we have more labels now. The following example shows the annotation used in the training data.

```
...<V>yemikahEdut</V>
<N>yemrmr</N>
<N>tegbarat</N>
<V>yatekorut</V>
<N>bemgb</N> <N>sebl</N>
...
```

Each word is enclosed in an XML tag that denotes its POS tag. These tags are directly used for the training of the sequence-labeling task. No additional reduction process is carried out.

The set of features that are used for training are composed of lexical features, morphological features, dictionary features, the previous two POS tags, and character bi-grams. We used a window of five words centered at the current word. The lexical features consist of the current word, the two words to the left and to the right of the current word. Morphological features are generated by extracting a segment of length one to four characters long from the beginning and end of the word. These segments are first checked against a precompiled list of valid prefix and suffix morphemes of the language. If the segment is a valid morpheme then an appropriate feature will be generated. Otherwise the null prefix or suffix feature will be generated to indicate the absence of an affix. The dictionary is used to generate a binary feature for a word based on the POS tag found in the dictionary. In other words, if the word is found in the dictionary, its POS tag will be used as a feature. For each word, a set of character bi-grams has been generated and each character bigram is used as a feature. Finally, the last two POS tags are also used as a feature.

8 Results

We conducted a 5-fold cross-validation experiment. In each run, one article is used as a test dataset and the remaining four articles are used for training. The results reported in the sections below are the average of these five runs. On average 80% of the words in the test files are unknown words. Most of the unknown words (on average 60%) are nouns.

8.1 Segmentation Result

As mentioned in Section 7.1, four sets of features, i.e., character features, morphological features, dictionary features, and previous label, are used for the segmentation task. Table 2 shows results for some combinations of these features. The results without the previous label feature are also shown (*Without Prev. Label*).

The simple character features are highly informative features, as can be seen in Table 2 (Row 1). Using only these features, the system with previous label feature already achieved an accuracy of 0.819. The dictionary feature improved the result by 2% whereas the morphological features brought minor improvements. As more features are added the variation between the different runs increases slightly. Performace significantly decreases when we omit the previous label feature as it is shown in *Without Prev. Label* column.

8.2 POS Tagging Results

Table 3 shows the word-based evaluation results of the POS tagging experiment. The baseline (Row 1) means assigning all the words the most frequently occurring POS tag, i.e., N (noun). The result obtained using only lexical features (Row 2) is better than the baseline. Adding morphological features improves the result almost by the same amount (Row 3). Incorporation of the dictionary feature, however, has brought only slight improvement. The addition of bi-gram features improved the result by 3%.

As mentioned before, it is not possible to compare the results, i.e. 74% accuracy (With Prev. Label), with other state of the art POS taggers since our data is very small compared to the data used by other POS taggers. It is also difficult to claim with absolute certainty as to the applicability of the technique we have applied. However, given the fact that 80% of the test instances are unseen instances, an accuracy of 74% is an acceptable result. This claim receives further support when we look at the results reported for unknown word guessing methods in other POS tagging experiments (Nakagawa et al., 2001). As we add more features, the system shows less variation among the different folds. As with segmentation task, the omission of the previous label feature decreases performace. The system with only lexical features and without previous label feature has the same performace as the baseline system.

8.3 Error Analysis

The results of both the segmentation and POS tagging tasks show that they are not perfect. An examination of the output of these systems shows certain patterns of errors. In case of the segmentation task, most of the words that are incorrectly segmented have the same beginning or ending charac-

	With Prev. Label		Without Prev. Label	
Features	accuracy	stddev	accuracy	stddev
Char.	0.819	0.7	0.661	4.7
Char.+Dict.	0.837	1.6	0.671	4.1
Char.+Dict.+Morph.	0.841	1.7	0.701	3.9

Table 2: Segmentation Results

	With Prev. Label		Without Prev. Label	
Features	accuracy	stddev	accuracy	stddev
Baseline	0.513	6.4	–	–
Lex.	0.613	5.3	0.513	6.4
Lex.+Morph.	0.700	5.0	0.688	5.2
Lex.+Morph.+Dict.	0.713	4.3	0.674	5.6
Lex.+Morph.+Dict.+Bigram	0.748	4.3	0.720	2.9

Table 3: Word-based evaluation results of POS tagging

ters as words with affix morphemes. Increasing the size of the lexical resources, such as the dictionary, can help the system in distinguishing between words that have affixes from those that do not.

The POS tagging system, on the other hand, has difficulties in distinguishing between nouns and other POS tags. This in turn shows how similar nouns are to words in other POS tags morphologically, since our experiment relies heavily on morphological features. This is not particularly surprising given that most Amharic affixes are shared among nouns and words in other POS tags. In Amharic, if a noun phrase contains only the head noun, most noun affixes, such as prepositions, definite article, and case marker appear on the head noun. If, on the other hand, a noun phrase contains prenominal constituents such as adjectives, numerals, and other nouns, then the above noun affixes appear on prenominal constituents, thereby blurring the morphological distinction between the nouns and other constituents. Furthermore, similar sets of morphemes are used for prepositions and subordinate conjunctions, which again obscures the distinction among the nouns and verbs. This, together with the fact that nouns are the dominant POS category in the data, resulted in most words being missclassified as nouns.

In general, we believe that the above problems can be alleviated by making more training data available to the system, which will enable us to determine improved parameters for both segmentation and POS tagging models.

9 Concluding Remarks

In this paper, we provided preliminary results of the application of CRFs for Amharic word segmentation and POS tagging tasks. Several features were examined for these tasks. Character features were found to be useful for the segmentation task whereas morphological and lexical features significantly improve the results of the POS tagging task. Dictionary-based features contribute more to the segmentation task than to the POS tagging task. In both experiments, omition of previous label feature hurts performance.

Although the size of the data limits the scope of the claims that can be made on the basis of the results, the results are good especially when we look at them from the perspective of the results achieved in unknown word recognition methods of POS tagging experiments. These results could be achieved since CRFs allow us to integrate several overlapping features thereby enabling optimum utilization of the available information.

In general, the paper dealt with a restricted aspect of the morphological analysis of Amharic, i.e., Amharic word segmentation and POS tagging. Furthermore, these tasks were carried out relatively independently. Future work should explore how these tasks could be integrated into a single system that

allows for fine-grained POS tagging of Amharic words. Parallel to this, resource development needs to be given due attention. As mentioned, the lack of adequate resources such as a large POS annotated corpus imposes restrictions on the kind of methods that can be applied. Therefore, the development of a standard Amharic POS tagset and annotation of a reasonably sized corpus should be given priority.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

References

Nega Alemayehu. 1999. *Development of stemming algorithm for Amharic text retrieval*. PhD Thesis, University of Sheffield.

Atelach Alemu. 2002. *Automatic Sentence Parsing for Amharic Text: An Experiment using Probabilistic Context Free Grammars*. Master Thesis, Addis Ababa University.

Amsalu Aklilu. 1987. *Amharic-English Dictionary*. Kuraz Publishing Agency.

Roy Bar-Haim, Khalil Simaan and Yoad Winter. 2004. Part-of-Speech Tagging for Hebrew and Other Semitic Languages. Technical Report.

Abiyot Bayou. 2000. *Developing automatic word parser for Amharic verbs and their derivation*. Master Thesis, Addis Ababa University.

W. Cohen. 2004. *Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data.* `http://minorthird.sourceforge.net`

Peter T. Daniels, 1997. Script of Semitic Languages in: Robert Hetzron, editor,*Proceedings of the Corpus Linguistics.* 16–45.

Mona Diab, Kadri Hacioglu and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From row text to base phrase chunks. In Daniel Marku, Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short papers*, pages 149–152, Boston, Massachusetts, USA, May 2–May 7. Association for Computational Linguistics

Sisay Fissaha Adafre and Johann Haller. 2003a. Amharic verb lexicon in the context of machine translation. *Traitement Automatique des Langues Naturelles* 2:183–192

Sisay Fissaha Adafre and Johann Haller. 2003b. Application of corpus-based techniques to Amharic texts. *Machine Translation for Semitic languages* MT Summit IX Workshop, New Orleans

Mesfin Getachew. 2001. *Automatic part of speech tagging for Amharic language: An experiment using stochastic HMM*. Master Thesis, Addis Ababa University.

Wolf Leslau. 1995. *Reference Grammar of Amharic*. Otto Harrassowitz, Wiesbaden.

A. McCallum and W. Li. 2003. Early results for Named Entity Recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the 7th CoNLL.*

Tetsuji Nakagawa, Taku Kudo and Yuji Matsumoto. 2001. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. NLPRS pages 325-331, Boston, Massachusetts, USA, May 2–May 7. `http://www.afnlp.org/nlprs2001/pdf/0053-01.pdf`

J. Lafferty, F. Pereira and A. McCallum. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning.*

F. Sha and F. Pereira. 2004. Shallow parsing with conditional random fields. *Proceedings of Human Language Technology-NAACL.*

Khoja S., Garside R., and Knowles G. 2001. A Tagset for the Morphosyntactic Tagging of Arabic *Proceedings of the Corpus Linguistics.* Lancaster University (UK), Volume 13 - Special issue, 341.

Leech G. Wilson. 1996. *Recommendations for the Morphosyntactic Annotation of Corpora.* EAGLES Report EAG-TCWG-MAC/R, `http://www.ilc.pi.cnr.it/EAGLES96/annotate/annotate.html`

Daniel Yacob. 1996. *System for Ethiopic Representation in ASCII.* `http://www.abyssiniagateway.net/fidel`

Baye Yimam. 1999. Root. *Ethiopian Journal of Language Studies*, 9:56–88.

Baye Yimam. 1986. *Yamara Swasw* E.M.P.D.A, Addis Ababa.