ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/237553785

Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges

Article · January 2006

CITATIONS
28

READS
2,623

2 authors, including:

Girma A. Demeke

11 PUBLICATIONS **79** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Storch, A. (2011). Secret Manipulations: Language and Context in Africa. Oxford: Oxford University Press. View project

All content following this page was uploaded by Girma A. Demeke on 06 April 2016.

The user has requested enhancement of the downloaded file.

Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges*

Girma Awgichew Demeke
&
Mesfin Getachew

Abstract

Since September 2005, the Ethiopian Languages Research Center of Addis Ababa University has been engaged in a project called "The Annotation of Amharic News Documents". The project was meant to tag manually each Amharic word in its context with the most appropriate parts-of-speech.

This paper presents the POS tagset developed for annotating the news documents, the problems encountered in the process of tagging the news documents and the procedures followed to manually tag them. The major output of the work contains 1065 Amharic news documents (that constitute 210,000 prosodic words) annotated manually with part-of-speeches and a new tagset for the language derived from the 1065 news item.

The outcome of the POS tagging project is assumed to have great contribution for future works in natural language processing of Amharic, including the development of probabilistic part-of-speech taggers (a software which uses a lexicon as a component for automatically assigning words with appropriate part-of-speech and a central component for higher level NLP tools such as parsers), a noun-phrase chunker (a software tool

* We would like to thank Walta Information Center, Ato Moges Delelegn for facilitating (along with Dr. Lars Asker of the University of Stockholm) the funding for the annotation project and the annotators for annotating the documents. Our especial thanks, however, goes to Dr. Asker for bringing the project to the Center, for arranging the funding and providing comments and every technical assistance through out the whole period of undertaking the project. Thanks also for .the participant of the "Workshop on Lexicography and Related Fields" organized by the Ethiopian Languages Research Center on April 7 & 8, 2006 for their valuable comments and suggestions.

that seeks to identify noun phrases in a text) and for works in speech synthesis, speech recognition, information retrieval, word sense disambiguation, corpus analysis and computational lexicography of Amharic.

1. Introduction

In this paper we present a recently completed project work by the Ethiopian Languages Research Center that deals with the parts-of-speech (POS) tagging of Amharic news items. The project was conducted since September 2005 for four months.

POS tagging is the process of assigning a POS or other lexical class marker to each word in a corpus (Jurafsky 2005). The project was initiated or stems from understanding the need for lack of basic Amharic resources that enable the construction of new resources by researchers (Alemu 2005, Getachew 2001). One such basic resource is a large corpus in the language annotated with POS information. In this, the prime objective of the project was to manually tag the 210,000 prosodic words that occur in the 1065 Amharic news documents with appropriate POSs or morpho-syntactic categories.

The news documents were donated by Walta Information Center and pre-processed by Dr. Lars Asker and Atelach Alemu of the University of Stockholm and provided in electronic copy to the Center by Dr. Lars Asker. Details of the pre-processing are available in Alamu (2005). Nine people, most of them from the Center, were involved in the actual manual tagging of the 1065 news documents. Besides, one technical assistant and four other administrative support staff were also involved at various levels during the project.

As it was unlikely to use the tagset for English by directly projecting English tags, a new POS tagset, derived from the 210,000 prosodic words of the Amharic news documents, was developed for use to tag the news documents mentioned. Tagsets of other languages (English, Korean, Arabic, and French), the nature of Amharic language itself and other constraints (like available fund, qualification of the annotators and time

assigned for undertaking the project) were carefully considered while identifying the tags for inclusion in the tagset.

The outcome of the POS tagging project is assumed to have great contribution for future works in Natural Language Processing of Amharic including the development of probabilistic part-of-speech taggers (a software which uses a lexicon as a component for automatically assigning words with appropriate part- of-speech and a central component for higher level NLP tools such as parsers), a noun-phrase chunker (a software tool in computational linguistics that seeks to identify noun phrases in a text) and for works in speech synthesis, speech recognition, information retrieval, word sense disambiguation, corpus analysis (cf.http://www.dcs.shef.ac.uk/~kiffer/teaching/tagging.pdf) and for computa- tional lexicography of Amharic.

The paper is organized as follows: Following this introductory section, section 2 presents a review of the Amharic word classes. In section 3, we discuss the parts-of-speech tagset developed for annotating the documents and problems in tagging the news documents. Section 4 concludes the paper.

2. Amharic parts-of-speech

Although Amharic is one of the most studied languages of Ethiopia, there is no consensus as to how many POS there are for this language. One of the most cited classical work on Amharic grammar, written in Amharic, i.e. Mersehazen (1935 E.C), put the word classes of this language as eight. These are preposition, noun, conjunction, interjection, verb, adjective, pronoun and adverb. In a recent work, Baye (1987 E.C.) suggests that Amharic has only five word classes. He does this leaving out interjection from the inventory and putting together prepositions and conjunctions in one class and considering pronouns as a sub-class of nouns. Baye's reduction of Mersehazen's classification seems based on the role of words in syntax, i.e. considering words that have clear role in Amharic sentential grammar.

In syntax, since pronouns act like nouns, categorizing pronouns under the class of nouns seems appropriate. Moreover, since most prepositions in Amharic function also as conjunctions putting these two into one word class again seems logical. Interjections are words that function beyond syntax. Leaving out this word class from the inventory of word classes, therefore, can also be justified if the focus is only on syntactic words.

However, even from syntactic point of view, Baye's classification cannot be considered as a refined and complete (cf. Demeke forthcoming). For instance, adjectives in Amharic, in fact, in other Semitic languages too, can be categorized as a sub-class of nouns. Adverbs too can be categorized as a sub-class of nouns, as is also the case in the traditional treatment of these word classes in Arabic. According to Demeke (ibid) Amharic has only four basic word classes. These are nominals, verbs, adpositions-conjunctions and interjections.

As is the case in any language, the above four classes can be sub-divided into sub-classes. The subclasses in turn can be divided into mini-sub-classes and the mini-sub-classes may further be divided into other mini-subclasses. Such division may go also further, depending on the level and aim of the investigation. In the following section, we discuss which classes and subclasses are considered for the annotation of the news items.

3. The Current Work

As mentioned earlier the aim of the project was to manually annotate 210,000 prosodic words found in 1065 Amharic news documents. The news documents were obtained from Walta Information Center, a private news agency located in Addis Ababa, Ethiopia, that makes daily news in Amharic and English through their website (http://www.waltainfo.com). The news items, as available in the website, are represented using Ethiopic script (Fidel) and were archived according to Ethiopian calendar.

A number of works have been done to bring the news texts obtained from Walta Information Center to a shape useful for tagging and making it

finally available in a form easier for users (mainly researchers in Amharic natural language processing). Putting aside the pre-processing task presented in detail in Alemu (2005) the major tasks of the current work focuses on identifying tags, building the tagset to be used for annotation, manually tag the news items with the tags identified and making available the final output on such websites as http://nlp.amharic.org and http://www.aau.edu.et using an Extensible Markup Language (XML).

Among the major tasks of the current work, the very challenging ones were the construction of the tagset for use to annotate the news items and to actually do the manual tagging. These tasks, during the project work, lead to many questions that require serious decisions at the outset. Among the questions that require serious decisions were how to get the tags for inclusion in the tagset, what information each tag should capture and the size of the tagset. We discuss these issues in the following sub-sections.

3.1 The Tagset

Needless to say, in listing the class of a certain word, the scope and the aim of the work in question has a crucial role. As a consequence, some sub-classes may be equally important to be listed along with the major classes whereas some others not.

In developing the POS tagset for the annotation of the news items, we have questioned which classes and sub-classes to be included in the final tagset. For our purpose, the basic POS identified are nouns (tagged N), pronouns (PRON), adjectives (ADJ), adverbs (ADV), verbs (V), prepositions (PREP), conjunctions (CONJ), and interjection (INT). Because punctuations should also be annotated, we have included a PUNC tag in the tagset. To give a room for tagging difficult or problematic words that the annotators may face, we have included a UNC tag, unclassified.

In the tagging project the aim was to give basic grammatical information for each "word". However, there is a restriction on the number of POS mostly due to financial constraint. The first suggestion was to build a

tagset that contains 20 POS tags. But, in the final procedure we ended up with a tagset having 30 POS tags.

Since the basic POS tagset identified are 11 in number (as discussed above) which is under the intended number of POS tags, initially we had planned to include further subclasses and some grammatical information such as number and gender for nouns and tense and aspect for verbs. However, when we looked at the documents, we have totally abandoned the idea of giving information on grammatical categories such as number, gender etc. We have only found a space to add four more POS tags ─ making a distinction between verbal nouns from the rest of nouns, dividing numerals into cardinal and ordinal and identifying three subclasses under verbs. These are auxiliaries, relative verbs and others.[1]

In the documents, not each word class or sub-class does always correspond to a prosodic word. Some word classes are realized cliticized with another word. For instance, adpositions and conjunctions are mostly realized as bound forms. Dividing phrasal prosodic words that contain more than one grammatical words into word classes (sub-classes) was problematic because in some cases the two word classes went phonological processes as in, for instance, **ባንድ** 'in one' and in some cases not, as in **በአንድ** 'in one'. Hence, as there is no consistency in writing through the documents, we had to abandon the idea of segmenting grammatical words that are realized as single prosodic words. Moreover, because of the bulk of the documents and the lack of proper expertise, doing the segmentation of such type of prosodic words could not be considered as an option. This left us with no alternative choice for considering the above suggested grammatical information in the POS tagger. But, rather it persuades us to differentiate between the mono-grammatical words from the other phrasal words. Hence, considering such facts the following POS tagset have been developed.

[1] The choice of these subclasses over the others is based on (1) their syntactic behavior, (2) their unique morphological shape and (3) their frequent occurrence in the texts.

Table 1: Tagset for annotating the 210,000 prosodic words

Basic Class	**Definition of the tag**	**Code of the Tag**
Noun	Verbal/ infinitival Noun, formed from any verb form such as active, passive, and repetitive, by attaching the prefix m(ä)-	VN
	Any noun including verbal noun attached with a preposition	NP
	Any noun including verbal noun attached with conjunction	NC
	Any noun including verbal noun with a proclitic preposition and an enclitic conjunction	NPC
	Any other noun; simple or derived	N
Pronoun	Pronoun attached with preposition	PRONP
	Pronoun attached with conjunction	PRONC
	Pronoun with a proclitic preposition and an enclitic conjunction	PRONPC
	Any other Pronoun	PRON
Verb	Auxiliary verb	AUX
	Relative verb	VREL
	Any Verb including relative verbs and auxiliaries attached with preposition	VP
	Any Verb including relative verbs and auxiliaries attached with conjunction	VC
	Any Verb including relative verbs and auxiliaries with a proclitic preposition and an enclitic conjunction	VPC
	Verb (all other)	V
Adjective	Adjective attached with preposition	ADJP
	Adjective attached with conjunctions	ADJC
	Adjective with a proclitic preposition	ADJPC

Basic Class	Definition of the tag	Code of the Tag
	and an enclitic conjunction	
	Any other Adjective	ADJ
Preposition	Preposition	PREP
Conjunction	Conjunction	CONJ
Adverb	Adverb	ADV
Numeral	Cardinal	NUMCR
	Ordinal	NUMOR
	Numeral (cardinal or ordinal) attached with preposition	NUMP
	Numeral (cardinal or ordinal) attached with conjunction	NUMC
	Numeral (cardinal or ordinal) with a proclitic preposition and an enclitic conjunction	NUMPC
Interjection	Interjections	INT
Punctuation	Punctuation	PUNC
Unclassified	Unclassified	UNC
Total		**30 tags**

Table 1 illustrates the tagset developed for the annotation of the news items. The table shows the final new tagset identified and used to manually annotate the 210,000 prosodic words of the Amharic news documents. It contains the definitions and abbreviations of each POS. As can be seen in the table, for instance, under noun we have five POS tags. From this list, one can see that only two types of nouns are identified. These are verbal nouns, VN, and other forms of nouns, N. This is basically because of the restriction set on the size of the POS tagset, as mentioned earlier. The distinction between VN versus N couldn't be maintained when such nouns appear with preposition, P, conjunction, C, or with both. Such distinction was not also mentioned in the other three basic POS tags under verbs and the two numerals identified in the tagset when they appear along with P, C or with both P&C.

3.2 Input text

3.2.1 The Initial text from Walta

The following is an excerption of the original electronic input text as it was donated by Walta Information Center.

በቦረና የኦህዴድ ተሃድሶ ውይይት ተጀመረ

መስከረም 07፣ 1994 በቦረና ዞንና 13 ወረዳዎች ለሚገኙ የመንግሥት ሠራተኞች የተዘጋጀ የኦህዴድ ተሃድሶ ውይይት ዛሬ መጀመሩን የዞኑ መስተዳድር ምክር ቤት አስታወቀ። የምክር ቤቱ ፀሐፊ አቶ መሐመድ ጅሎ እንደገለፁት ለአምስት ቀናት በሚቆየው በዚሁ ተሃድሶ የአብዮታዊ ዴሞክራሲያዊ ጥያቄ በኢትዮጵያ፣ የአብዮታዊ ዴሞክራሲ የልማት መርሆዎች፣ ስትራቴጂዎችና የሥርዓቱ አደጋዎች በሚሉ ርዕሶች ላይ ውይይት ይካሄዳል። የአመለካከትን ጥራት ለማምጣት በሚካሄደው የተሃድሶ ውይይት ከዞን መምሪያዎችና ከወረዳ ጽሕፈት ቤቶች የተውጣጡ ከ2 ሺ 500 በላይ የመንግሥት ሠራተኞች ይሳተፋሉ ተብሎ እንደሚጠበቅ ፀሐፊው ለዋልታ ኢንፎርሜሽን ማዕከል ገልፀዋል።

The original input text contains news items of the above kind produced by the Information Center from “Meskerm” to “Mizia” 1994 (E.C.). Each news document has two sections, title and body of the news, written using the Amharic Fidel. In the body there are information about the month, date and year in which the news in question was produced and information about the agency that produced the news. As mentioned earlier the total number of the news documents considered in the annotation project were 1065 that contain in total 210,000 prosodic words.

3.2.2 The modified input text

The modified or pre-processed input text for the intended work was a result of pre-processing done in Alemu (2005). It is presented to the Center in an XML format and has the following structure.

```
<?xml version="1.0"?>
<!DOCTYPE amnews94 >
  <!ELEMENT document (filename, title, dateline, body)>
  <!ELEMENT filename (#PCDATA)>
  <!ELEMENT title (fidel, sera)>
  <!ELEMENT fidel (#PCDATA)>
  <!ELEMENT sera (#PCDATA)>
  <!ELEMENT dateline EMPTY>
  <!ATTLIST dateline place CDATA #IMPLIED>
  <!ATTLIST dateline month CDATA #IMPLIED>
  <!ATTLIST dateline date CDATA #IMPLIED>
  <!ELEMENT body (fidel, sera)>
```

The set of rules above declare that the element to be described is amnews94, Amharic news text in the year 1994 (E.C.). Each Amharic news text or document has four data elements, namely file name, title, date and body whose values are parsed character data (represented by #PCDATA). The title and the body of each news text are represented both in Ethiopic Script (Fidel) and SERA (Yacob 1996). The date (represented as dateline) has attributes by the name place, month and date to indicate the place, month and year (and news agency) that the news produced respectively. The values of the dateline are all character data (represented as CDATA, the same as PCDATA. each of which are implied, i.e., values of the attributes are not required or may not be provided

The following is an excerption of a single Amharic news text from the modified input text supplied in electronic form.

```
<document>
 <filename>mes07a2.htm</filename>
  <title>
  <fidel>በቦረና የኦህዴድ ተሃድሶ ውይይት ተጀመረ</fidel>
  <sera>beborena yeohdEd tehadso wyyt tejemere</sera>
</title>
 <dateline place="negelE" month="meskerem" date="7/1994/(WIC)/" />
```

<body>
<fidel>በቦረና ዞንና 13 ወረዳዎች ለሚገኙ የመንግሥት ሠራተኞች የተዘጋጀ የኦህዴድ ተሃድሶ ውይይት ዛሬ መጀመሩን የዞኑ መስተዳድር ምክር ቤት አስታወቀ። የምክር ቤቱ ፀሐፊ አቶ መሐመድ ጅሎ እንደገለፁት ለአምስት ቀናት በሚቆየው በዚሁ ተሃድሶ የአብዮታዊ ዴሞክራሲያዊ ጥያቄ በኢትዮጵያ፣ የአብዮታዊ ዴሞክራሲ የልማት መርሆዎች፣ ስትራቴጂዎችና የሥርዓቱ አደጋዎች በሚሉ ርዕሶች ላይ ውይይት ይካሄዳል። የአመለካከትን ጥራት ለማምጣት በሚካሄደው የተሃድሶ ውይይት ከዞን መምሪያዎችና ከወረዳ ጽሕፈት ቤቶች የተውጣጡ ከ2 ሺ 500 በላይ የመንግሥት ሠራተኞች ይሳተፋሉ ተብሎ እንደሚጠበቅ ፀሐፊው ለዋልታ ኢንፎርሜሽን ማዕከል ገልፀዋል።</fidel>

<sera>beborena zonna 13 weredawoc lemigeNu yemeng`st `serateNoc yetezegaje yeohdEd tehadso wyyt zarE mejemerun yezonu mestedadr mkr bEt astaweqe:: yemkr bEtu `SeHefi ato meHemed jlo Indegele`Sut leamst qenat bemiqoyew bezihu tehadso yeabyotawi dEmokrasiyawi TyaqE beityoPya, yeabyotawi dEmokrasi yelmat merhowoc, stratEjiwocna ye`sr`atu adegawoc bemilu r`Isoc lay wyyt ykahEdal:: yeamelekaketn Trat lemamTat bemikahEdew yetehadso wyyt kezon memriyawocna kewereda SHfet bEtoc yetewTaTu ke2 xi 500 belay yemeng`st `serateNoc ysatefalu teblo IndemiTebeq `SeHefiw lewalta informExn ma`Ikel gel`Sewal::</sera>

<copyright>Copyright 1998 - 2002 Walta Information Center</copyright>
</body>
</document>

3.4. The procedure for manual tagging

After the annotators, all with linguistics background, were identified, a brief instruction was given to them on how to annotate the news items. A brief lecture was also given about the nature of Amharic POS taking into consideration the data. Then sample news texts from the collection were distributed to the annotators. Based on the annotators work on the given sample texts discussions were made which finally resulted for improving the initial POS tagset.

News documents were then distributed to the annotators in hard copies after we reached on agreement on the POS tagset, and that the annotators became familiar with the tagset developed and the POS tagging process. Each annotator got different news documents. Then the annotators did the POS tagging for each news item, with pen on the hard copies given to them. Then the manually annotated documents (in hard copies) were checked at a glance (which sometimes resulted in returning to the annotators for revision) and given to typists for inserting the tags in the electronic version.

Then the tagged news documents were reviewed several times for errors in tag assignments. However, because of the bulk of the documents, shortage of funding and time constraints, we couldn't engage many linguists for review. Hence, we did not feel that the final annotated news documents are error free. Sample of the POS tagged news items are presented in the following section.

3.5.1 Output text

The following is an excerption of a single Amharic news text from the final output of the project. The only difference from the modified or pre-processed input text discussed earlier is that each word in the Latin representation is marked with appropriate tag in the form word <tag>. That is, the most probable tag is attached in angle bracket immediately after each word leaving one space after each prosodic word.

<document>
<filename> mes07a2.htm </filename>
<title>
<sera>
beborena <NP> yeohdEd <NP> tehadso <N> wyyt <N> tejemere <V> ::
<PUNC>
</sera>
</title>
<dateline place="negelE" month="meskerem" date="7/1994/(WIC)/" />
<body>

<fidel>በቦረና ዞንና 13 ወረዳዎች ለሚገኙ የመንግሥት ሠራተኞች የተዘጋጀ የኦህዴድ ተሃድሶ ውይይት ዛሬ መጀመሩን የዞኑ መስተዳድር ምክር ቤት አስታወቀ። የምክር ቤቱ ፀሐፊ አቶ መሐመድ ጅሎ እንደገለፁት ለአምስት ቀናት በሚቆየው በዚሁ ተሃድሶ የአብዮታዊ ዴሞክራሲያዊ ጥያቄ በኢትዮጵያ፣ የአብዮታዊ ዴሞክራሲ የልማት መርሆዎች፣ ስትራቴጂዎችና የሥርዓቱ አደጋዎች በሚሉ ርዕሶች ላይ ውይይት ይካሄዳል። የአመለካከትን ጥራት ለማምጣት በሚካሄደው የተሃድሶ ውይይት ከዞን መምሪያዎችና ከወረዳ ጽሕፈት ቤቶች የተውጣጡ ከ2 ሺ 500 በላይ የመንግሥት ሠራተኞች ይሳተፋሉ ተብሎ እንደሚጠበቅ ፀሐፊው ለዋልታ ኢንፎርሜሽን ማዕከል ገልፀዋል።</fidel>

<sera>
beborena <NP> zonna <N> 13 <NUMCR> weredawoc <N> lemigeNu <VP> yemengst <NP> serateNoc <N> yetezegaje <VREL> yeohdEd <NP> tehadso <N> wyyt <N> zarE <ADV> mejemerun <VN> yezonu <NP> mestedadr <N> mkr <N> bEt <N> astaweqe <V> :: <PUNC> yemkr <NP> bEtu <N> Sehefi <N> ato <ADJ> mehemed <N> jlo <N> IndegeleSut <VP> leamst <NUMP> qenat <N> bemiqoyew <VP> bezihu <PRONP> tehadso <N> yeabyotawi <ADJP> dEmokrasiyawi <ADJ> tyaqE <N> beityoPya <NP> , <PUNC> yeabyotawi <NP> dEmokrasi <N> yelmat <NP> merhowoc <N> , <PUNC> stratEjiwocna <NC> yesratu <NP> adegawoc <N> bemilu <NP> rIsoc <N> lay <PREP> wyyt <N> ykahEdal <V> :: <PUNC> yeamelekaketn <NP> trat <N> lemamTat <NP> bemikahEdew <VP> yetehadso <NP> wyyt <N> kezon <NP> memriyawocna <NC> kewereda <NP> shfet <N> bEtoc <N> yetewTaTu <VREL> ke2 xi 500 <NUMP> belay <NP> yemengst <NP> serateNoc <N> ysatefalu <V> teblo <V> IndemiTebeq <VP> Sehefiw <N> lewalta <NP> InformExn <N> maIkel <N> gelSewal <V> :: <PUNC>
</sera>

<copyright> copyright 1998 - 2002 Walta Information Center </copyright>
</body>
</document>

3.6 Challenges faced

As we have indicate in section 3 .1 one of the major challenges in tagging the Amharic news items in particular and Amharic texts in general is that word classes in the language may not necessarily correspond to prosodic words. Some prosodic words in the text contain more than one grammatical words, that belong to different word classes. On the other hand, for the already mentioned reason dividing the prosodic words into morphological segments (word classes) was not in the scope and aim of the project. As a result we introduced a number of special tags such as NP, NC, NPC, PRONP, PRONC, PRONPC and so on to address this issue, as we have seen in section 3.1.

The other problems have to do with frequently mis-tagged words. These include (1) Nouns that come with preposition having adverbial function, which are tagged as adverbs, (2) words such as **ዝም፣ ብድግ** etc. which are tagged differently by different annotators and (3) Relative verbs which are frequently tagged as adjectives.

4. Conclusion

In this paper we have described the work done by the Ethiopian Languages Research Center of Addis Ababa University on manual annotation of Amharic news items with appropriate part-of-speeches. We have presented the steps that have been taken, the tagset developed and the challenges faced to manually tag 1065 news items containing 210, 000 prosodic words. Among the challenges discussed are determining the tagset size, how to deal with prosodic words that are at phrase levels and to actually do the manual (or hand) tagging, an intellectual activity which is expensive and difficult.

The output, which will be available on different websites, although tiny, is believed to be a useful resource for developing probabilistic part-of-speech tagger, a noun-phrase chunker and for works in speech synthesis, speech recognition, information retrieval, word sense disambiguation, corpus analysis and lexicography of Amharic.

The final output is the first version and is not error free. Updated versions will be released incorporating necessary changes on continuous bases. Since the text used in the annotation is tiny and not representative of the language, the tagset obtained cannot be representative. Thus, as a continuation of this project, the following are suggested as future tasks.

- Develop Treebanks for Amharic (and other local languages);
- Build a tagset for Amharic (and other local languages) using the Treebanks;
- Build lexicons that could be used as a components for probabilistic POS taggers; and
- Develop (probabilistic. part-of-speech taggers, which could be used to tag Amharic news texts automatically using bootstrapping technique.

References

Alemu, A. and Asker, L. 2005. "Web Mining for an Amharic -English Bilingual Corpus", in Proceedings of the 1st International Conference on Web Information Systems and Technologies (WEBIST 2005), Miami.

Baye Yimam 1987. E.*C. yamariñña säwasiw (Amharic Grammar)*. Addis Ababa: EMPDA.

Demeke, Girma A. (forthcoming). Amharic Word Classes. WCAL 5, August 2006, Addis Ababa University.

Jurafsky, D. and James, H. 2000. *Speech and Language Processing*. Prentice Hall:

Mersehazen Wolde Kirkos. 1935 E.C. *Amharic Grammar* (text in Amharic). Addis Ababa: Artistic Priniting Press.

Yacob, D. (1996). System for Ethiopic Representation in ASCII (SERA). http://www.abyssiniacybergateway.net/fidel/.

Addresses of the authors:
Girma A. Demeke
Ethiopian Languages Research Center, Director
Addis Ababa University
Email: girmaad@gmail.com

&

Mesfin Getachew
Faculty of Informatics, Department of Information Science
Addis Ababa University
Email: mesgetachew@yahoo.com

View publication stats