Tagging and Verifying an Amharic News Corpus

Björn Gambäck

Norwegian University of Science and Technology
Trondheim, Norway
gamback@idi.ntnu.no

Abstract

The paper describes work on verifying, correcting and retagging a corpus of Amharic news texts. A total of 8715 Amharic news articles had previously been collected from a web site, and part of the corpus (1065 articles; 210,000 words) then morphologically analysed and manually part-of-speech tagged. The tagged corpus has been used as the basis for testing the application to Amharic of machine learning techniques and tools developed for other languages. This process made it possible to spot several errors and inconsistencies in the corpus which has been iteratively refined, cleaned, normalised, split into folds, and partially re-tagged by both automatic and manual means.

1. Introduction

There is a major shortage of language processing tools and resources for (almost all) African languages. This paper focuses on Amharic, the primary language of Ethiopia, and on the correction and processing of a tagged Amharic text corpus, taking as the starting-point a set of Amharic news articles collected at Stockholm University from an Ethiopian web news archive, and then morphologically analysed and manually part-of-speech tagged at Addis Ababa University.

Amharic is the second largest Semitic language: speakers of Arabic count in hundreds of millions, of Amharic in tens of millions, and of Hebrew and Tigrinya in millions. Ethiopia is divided into nine regions, each having its own official language, but Amharic is used as the *lingua franca* at the national level, so the number of second language speakers is fairly high. Giving a figure for the size of the Amharic speaker body is not easy, but based on the latest Ethiopian census carried out in 2007 (CSA, 2010), estimates of the current population of Ethiopia (CIA, 2012), and approximations of the percentage of Ethiopians speaking Amharic given the previous census in 1994 (Hudson, 1999), it is reasonable to assume that about 30 million persons speak it as first language and more than 10 million as second language.

Amharic uses a unique script (shared with Tigrinya), which in contrast to Arabic and Hebrew is written from left to right. The script is commonly known as "Ethiopic script", "Ge'ez" or *fidel* (lit. "alphabet" in Amharic) and is basically syllabic with most of its 275 characters representing consonant-vowel combinations. The language is quite diversified both when spoken and written, and has, for example, no standard spelling for compounds and loan-words. It has a complex morphology, where nouns (and adjectives) are inflected for gender, number, definiteness, and case. Definite markers and conjunctions are suffixed to the nouns, while prepositions are prefixed. Like other Semitic languages, the verbal morphology is rich and based on triconsonantal roots.

Despite the large number of speakers, there have been few efforts to create language processing tools for Amharic. A deterrent to progress was lack of standardisation: an international standard for Ethiopic script was agreed on only in 1998 and incorporated into Unicode in 2000. Several representation formats for the script were used before that, thwarting language processing and electronic publication in Amharic. As an effect, almost all work on Amharic language processing has taken place after the millenium shift.

Another major deterrent to progress in Amharic language processing has been the lack of large-scale resources such as corpora and tools. Thus, for example, the best result reported by an Amharic part-of-speech tagger before the availability of the corpus discussed in this paper was by Adafre (2005). That work suffered from only having access to a 1,000 word training corpus, resulting in a word error rate of over 25%. As we shall see (in Section 5.), that can be improved to figures below 10% using a 200k word corpus; a number which still is high though, when compared to better-resourced language, for which WER of 2–4% is common.

Language processing for Amharic has in fact taken two major steps forward in recent years, both through the creation of a reasonably-sized tagged corpus and through the appearance of HornMorpho (Gasser, 2009), the most complete morphological processing tool for Amharic (and Tigrinya) to date. HornMorpho uses a finite-state approach to allow for both analysis and generation of nouns (and adjectives which are regarded as nouns) and verbs. Gasser (2011) attempts a small evaluation of HornMorpho's performance on 200 randomly selected words each for the two "wordclasses", reporting that about 95.5% of the noun/adjectives, and 99% of the verbs received an analysis (i.e., all legal combinations of roots and grammar structures for them could be found).

The Amharic news corpus in the present paper had previously been extracted from the web and manually part-of-speech tagged, as described in Section 2. The core of the paper is the normalisation and clean-up measures that had to be performed after the manual tagging (Section 3.) and the re-tagging and splitting of the corpus into folds (Section 4.). The corpus has been tested by the application to Amharic of several machine learning techniques for part-of-speech tagging, discussed in Section 5. This process enabled the spotting of errors and inconsistencies in the corpus, which was subsequently refined both automatically and manually.

2. Creating the Corpus

Corpora are commonly being distinguished by being untagged or tagged (that is, marked up with tags such as part-of-speech or sentence structure, etc.), as well as by being balanced or domain-specific. While a balanced corpus would provide a wide selection of different types of texts, the corpus discussed in this paper is made up of texts from the news domain only: all texts used in the corpus come from the web archives of Walta Information Center (`www.waltainfo.com`), a private Addis Ababa-based news and information service providing daily Ethiopia-related news coverage in Amharic and English.

All Amharic news items (8715 in total) from the start of the service in March 2001 to December 2004 were downloaded from the Walta web archive using a web-crawler and stored in an XML structure by staff at the Department of Computer and System Sciences, Stockholm University (Argaw and Asker, 2005).

Due to the above-mentioned lack of standard representation, a variety of ways had been used to encode the *fidel* in the Amharic texts. In order to have a unified representation of the corpus and to simplify further analysis, the Stockholm University staff transliterated all texts into SERA, "System for Ethiopic Representation in ASCII" (Yacob, 1997), a convention for transcription of Ethiopic into 7-bit ASCII, and then back to a common Unicode compatible font (Ethiopia Jiret). Both the Ethiopic and the SERA transliterated form are stored in the XML structure of the corpus, under different fields. The full, untagged corpus contains about 1.7 million words after preprocessing (Argaw and Asker, 2005).

A portion of the untagged corpus was selected for manual tagging. This part of the corpus consisted of 1065 news texts from September 11, 2001 to May 8, 2002 (i.e., the first eight months of the Ethiopian year 1994; the Ethiopian calendar runs approximately seven years and eight months behind the Gregorian). In total 207,315 words in the Unicode version (*fidel*), and 207,291 words in the SERA-transcribed version.

The texts were tagged by staff at ELRC, the Ethiopian Languages Research Center at Addis Ababa University, using a tag set developed at ELRC and described by Demeke and Getachew (2006). The tagset is made up of thirty classes based on type of word only: the tags contain no information on grammatical categories (such as number, gender, tense, and aspect). Table 1 contains short explanations of the different classes in the column labeled "Description". The tags used in the manual tagging are the ones in the column called 'ELRC' in the table (which is an adapted version of the word-class table given by Demeke and Getachew). The annotation was carried out by nine trained linguists, who wrote the proposed tags with pen on hard copies. The hand-written tags were later typed out and digitalized by non-linguists.

The tagged corpus is available at `nlp.amharic.org`.

3. Cleaning the Corpus

Unfortunately, the corpus available on the net contains quite a few errors and tagging inconsistencies: several portions contain inconsistent manual tagging, in addition to the inter-annotator disagreement which can be expected in any manual tagging endeavor.

3.1. Possible Error Sources

Obviously, the tagging procedure introduced several possible error sources, with nine persons doing the tagging and others inserting the hand-written tags into the electronic version of the corpus. Many errors are also due to lack of resources: lack of trained personnel and time/funding for the tagging meant that each section of the corpus was only tagged by one person, while lack of computational resources meant that typists (non-linguists) rather than the linguists themselves entered the tags into the digital version of the corpus —- leaving room for misreadings and misinterpretations of the hand-written tags by the typists. Finally, the transliteration of the texts written in Ethiopic script into the SERA (ASCII) version actually also added some errors, thus, for example, a few non-ASCII characters still remain in the ASCII part of the published corpus.

In its present state, the corpus is still useful, but a main aim of the present work was to improve its quality and to "clean" it. To this end, many non-tagged items have been tagged. In contrast, some items in the corpus contained double tags, which have been removed.

3.2. Multi-Word Expressions

The on-line tagged corpus contains several headlines of the news texts tagged as one multi-word unit, assigned the end-of-sentence punctuation tag (`<PUNC>`), while some headlines have no tag at all. For those cases all the words of the headlines were re-tagged separately.

In contrast, the corpus also contains several "true" multi-word expressions (MWE) that have been assigned a single tag by the annotators. However, since the words still can be written separately, that has introduced a source of error, with the non-final words of the MWEs having no tag directly assigned to them. Reflecting the segmentation of the original Amharic text, all white-spaces were removed from the SERA-transcribed version, merging multi-word units with a single tag into one-word units. The alternative would have been to assign new tags to each of the words in a collocation.

Both approaches have pros and cons. On one hand, tagging each part of an MWE separately increases the size of the corpus (measured in words), and potentially reduces ambiguity by creating fewer lexical units. On the other hand, if the human annotator had assigned a single tag to the entire MWE, it makes sense to also attach a single tag to it, reflecting the choice of the linguist: It is important to correct pure mistakes in the human annotation, but not to interfere with *conscious* decisions taken by the annotators. Furthermore, attaching tags to different parts of a collocation is problematic due to the lack of spelling standards for compounds in Amharic: given that a unit AB can be written both as $A\ B$ and as AB in the source texts, it was deemed best to aim for consistency and remove any white-spaces inside the collocations (i.e., in effect enforcing the AB option).

3.3. Inconsistencies

Items such as `'`, `/`, etc., had been assigned a range of different tags, but have now been consistently re-tagged as punctuation (`<PUNC>`). Consistent tags have also been added to word-initial and word-final hyphens, by changing them into

Class	Tag description	ELRC	BASIC	SISAY
	Nouns			
1	Any basic or derived noun not matching classes 2–5	N	N	N
2	Verbal/infinitival noun, formed from any verb form	VN	N	N
3	Noun attached with a preposition	NP	N	N
4	Noun attached with a conjunction	NC	N	N
5	Noun with a preposition and a conjunction	NPC	N	N
	Pronouns			
6	Any pronoun not matching classes 7–9	PRON	PRON	**N**
7	Pronoun attached with a preposition	PRONP	PRON	**N**
8	Pronoun attached with a conjunction	PRONC	PRON	**N**
9	Pronoun with a preposition and a conjunction	PRONPC	PRON	**N**
	Verbs			
10	Any verb not matching classes 11–15	V	V	V
11	Auxiliary verb	AUX	**V**	AUX
12	Relative verb	VREL	V	V
13	Verb attached with a preposition	VP	V	V
14	Verb attached with a conjunction	VC	V	V
15	Verb with a preposition and a conjunction	VPC	V	V
	Adjectives			
16	Any adjective not matching classes 17–19	ADJ	ADJ	AJ
17	Adjective attached with a preposition	ADJP	ADJ	AJ
18	Adjective attached with a conjunction	ADJC	ADJ	AJ
19	Adjective with a preposition and a conjunction	ADJPC	ADJ	AJ
	Numerals			
20	Cardinal numeral not matching classes 22–24	NUMCR	NUM	NU
21	Ordinal numeral not matching classes 22–24	NUMOR	NUM	NU
22	Numeral attached with a preposition	NUMP	NUM	NU
23	Numeral attached with a conjunction	NUMC	NUM	NU
24	Numeral with a preposition and a conjunction	NUMPC	NUM	NU
	Others			
25	Preposition	PREP	PREP	AP
26	Conjunction	CONJ	CONJ	**AP**
27	Adverb	ADV	ADV	AV
28	Interjection	INT	INT	I
29	Punctuation	PUNC	PUNC	PU
30	Unclassified	UNC	UNC	R

Table 1: The three tagsets with tag descriptions (bold-face tags mark genuine differences)

two tokens, and making the following replacements:

$$-\textit{word}\ \texttt{<TAG>} \quad\Rightarrow\quad -\ \texttt{<PUNC>}\ \textit{word}\ \texttt{<TAG>} \tag{1}$$

$$\textit{word}-\ \texttt{<TAG>} \quad\Rightarrow\quad \textit{word}\ \texttt{<TAG}'\texttt{>}\ -\ \texttt{<PUNC>} \tag{2}$$

For the word-final case (2) this sometimes meant that a new tag (<TAG′>) had to be introduced for a word, namely in those cases when the originally assigned tag was <PUNC> (i.e., the tag relating to the -), rather than the tag relating to *word*. In the same fashion, a couple of cases of - tagged as <N> or <NP> were changed to punctuation, <PUNC>.

Making these changes is consistent with about half of the previous manual tagging of the corpus.

3.4. Tagging Errors and Misspellings

Furthermore, some direct tagging errors and misspellings have been corrected. Those include tag misspellings (or typos), such as <PUNC instead of <PUNC>. The published corpus also contains 14 cases similar to

$$\texttt{leityoPya <NP>rEdiyo <N>} \tag{3}$$

that is, with >rEdiyo written without a space. Those have all been corrected, as well as: two occurrences of meaningless] in the text, three occurrences of superfluous $ characters, three occurrences of \ characters, and one occurrence of a ... sequence.

3.5. Known Remaining Problems in the Corpus

Both time expressions and numbers in the corpus suffer from not having been consistently tagged at all, but just removing some of them can hardly be the way to handle that, so those had to be left as they were. In addition, many words have been transcribed into SERA in several versions, with only the cases differing. However, this is also difficult to account for since the SERA notation in general lets upper and lower cases of the English alphabet represent different symbols in *fidel* (Ethiopic script). Thus those were also left unchanged.

After corrections the SERA-transcribed version of the corpus contains 200,863 tagged words (compared to 207,291 words with 200,533 tags in the original corpus). 33,408 are unique wordforms, with 39,921 possible word-tag combinations (32,556 resp. 40,510 in the original corpus). 86% of the wordforms (28,731) have only one possible tag assignment, that is, are unambiguous. 3,533 have two possible tags, 744 have three, and 400 have four or more, including two words (*beteleym* and *yahl*) that have been given eleven possible tag assignments each. In the original corpus, 82% (26,844 wordforms) were unambiguous.

3.6. Alternative Cleaning Methods

In addition to the cleaning procedure described here, the corpus has also been partially cleaned in two other independent efforts. Hence Tachbelie (2010) primarily targeted inconsistencies in the annotations of collocations. She assigned separate tags to all the words in an MWE, rather than merging the words in the MWE and assigning it a single tag, as in the approach taken in Section 3.2. After cleaning, Tachbelie's version of the corpus contained 205,355 tokens.

Gebre (2010) carried out a more thorough cleaning of the corpus, in many ways following a strategy similar to the one of the present paper, but with the addition of treating collocations similarly to Tachbelie (2010). After cleaning, his version of the corpus contained 206,929 tokens. Quite importantly, the number of ambiguous tags was drastically reduced: in the original corpus only 38% of the tokens were unambiguous, but after removing tagging inconsistencies, this increased to 74% (compared to 82% resp. 86% above).

4. Re-Tagging and Splitting the Corpus

The manual tagging of the corpus utilized a 30-class tag set described by Demeke and Getachew (2006). The new, corrected corpus has been marked up with three different tagsets, all shown in Table 1.

4.1. Three Tagsets

Firstly, the full, original 30-tag set developed at the Ethiopian Languages Research Center. This version of the corpus will hereinafter be referred to as 'ELRC'. It differs from the published corpus in way of the corrections described in the previous section.

Secondly, the corpus was mapped to 11 basic tags also given by Demeke and Getachew (2006). This set consists of ten word classes: Noun, Pronoun, Verb, Adjective, Preposition, Conjunction, Adverb, Numeral, Interjection, and Punctuation, plus one tag for problematic words (unclear: <UNC>). This tagset will be called 'BASIC' below.

The main differences between the twose tagsets pertain to the treatment of prepositions and conjunctions: in 'ELRC' there are specific classes for, e.g., <PRONP>, <PRONC>, and <PRONPC>, i.e., for pronouns attached with preposition, conjunction, and both proclitic preposition and enclitic conjunction (similar classes occur for nouns, verbs, adjectives, and numerals). In addition, numerals are divided into cardinals and ordinals, verbal nouns are separated from other nouns, while auxiliaries and relative verbs are distinguished from other verbs. Hence the 'ELRC' tagset is made up of thirty subclasses of the eleven 'BASIC' classes.

Baseline	ELRC	BASIC	SISAY
Most frequent tag overall	35.50	58.26	59.61
Most frequent tag for word	79.64	83.05	83.10
Most likely tag	82.64	90.07	90,19

Table 2: Baselines for the three tagsets

Thirdly, for comparison reasons, the 'ELRC' tagset was also mapped to a tagset introduced by Adafre (2005), which will be referred to as 'SISAY' and is shown in the right-most column of Table 1. It consists of 10 different classes, including one for Residual (R) which was assumed to be equivalent to <UNC>. In addition, both <CONJ> and <PREP> were mapped to Adposition (AP) in Adafre's classification, and both <N> and <PRON> to N. The other mappings were straight-forward, except that the 'BASIC' tagset groups all verbs together, while Adafre (2005) kept Auxiliary (AUX) as its own class. The tags in bold-face in Table 1 indicate the major differences between this tagset and the 'BASIC' one. (Note that the verb classes 13–15 include auxiliaries attached with prepositions and/or conjunctions. It is unclear whether Adafre kept those together with the V class or with AUX; here they have been assumed to belong to the V class.)

Table 2 shows baselines for the tagsets. The "most frequent tag overall" baseline is the number of tokens tagged with the tag most frequent overall in the corpus (i.e., regular nouns; <N>) in relation to the total number of tokens (200,863). The "most frequent tag for word" baseline is the accuracy which would be obtained if labeling every word with the tag occurring most frequently with it in the corpus. The "most likely tag" baseline is the hardest to beat since it combines the other two, labeling known words with their most frequent tags and unknown words with <N>. It is fair to say that the higher the baseline, the easier it is to assign the tags of the tagset to a corpus: by simply guessing that a word is a standard noun, almost 6 out of 10 words would be correctly tagged using the 'SISAY' tagset or 'BASIC' tagsets. However, only 1 out of 3 words would be correctly tagged if applying the same strategy to the 'ELRC' tagset.

4.2. Splitting the Corpus into Folds

For evaluation of machine learning and statistical methods, it is common and useful to split a corpus into ten folds, each containing about 10% of the corpus. However, this can be done in several ways, for example, by taking the first 10th of the corpus as the first fold or by taking the folds to include every 10th word in the corpus (i.e., fold 1 consisting of words 1, 11, 21, etc.). The folds can also be of exactly equal size or be allowed to vary somewhat in size in order to preserve logical units (e.g., to keep complete sentences in the same fold). Furthermore they can be *stratified*, meaning that the distributions of different tags are equal over all folds.

The importance of how the folds are created points to one of the problems with n-fold cross validation: even though the results represent averages after n runs, the choice of the original folds to suit a particular machine learning or statistical strategy can make a major difference to the final result. For ease of straight-forward comparison between different studies, the same folds have to be used (cmf. the discussion in the next section, in particular the results of the MaxEnt tagger on different folds). We thus created a "standard" set of folds for the corpus.

The "standard" folds were created by chopping the corpus into 100 pieces, each of about 2000 words in sequence, while making sure that each piece contained full sentences (rather than cutting off the text in the middle of a sentence), and then merging sets of ten pieces into a fold. Thus the folds represent even splits over the corpus, to avoid tagging inconsistencies, but the sequences are still large enough to potentially make knowledge sources such as n-grams useful.

The resulting folds on average contain 20,086 tokens. Of those, 88.26% (17,727) are known, while 11.74% (2,359) are unknown, that is, tokens that are not in any of the other nine folds (if those were used for training, in a 10-fold evaluation fashion). Notably, the fraction of unknown words is about four times higher than in the English Wall Street Journal corpus (which, in contrast, is about six times larger).

5. On Tagging the Untagged Corpus

Having a tagged corpus is useful for many types of statistical and machine-learning based approaches to language processing. As an example application we will here look at automatic part-of-speech tagging, that is, the task of automatically assigning exactly those tags to the words that the human annotators assigned (or should have assigned). This could be a straight-forward way to extend the tagged corpus in itself, since the manually tagged portion of the corpus is only about 12% (1065 of 8715 news items). Clearly, tagging the remaining corpus would be useful, but as pointed out in Section 3.1., manually annotating the entire corpus would be an endeavour which would have to rely on human resources that are both scarce, expensive and inconsistent.

Part-of-speech tagging is a classification task, with the object of assigning lexical categories to words in a text. Within the computational linguistic community, part-of-speech tagging has been a fairly well-researched area. Most work so far has concentrated on English and on using supervised learning methods. The best results on the English Wall Street Journal corpus are now above 97%, using combinations of taggers: Spoustová et al. (2009) report achieving an accuracy of 97.43% by combining rule-based and statistically induced taggers. However, recently the focus has started to shift towards other languages and unsupervised methods.

The best reported figures for part-of-speech tagging for Arabic are comparable to those for English (Habash and Rambow, 2005; Mansour, 2008), while those for other Semitic languages are a bit lower: with 21 tags and a 36,000 word news text, Bar-Haim et al. (2008) report 86.9% accuracy on Hebrew. For Amharic, the best results reported before the availability of the corpus discussed here came from experiments by Adafre (2005) on using a Conditional Random Fields (CRF) tagger, an effort restricted by only having access to a news text corpus of 1,000 words. Using the 10-tag 'SISAY' tagset (see Table 1), the tagger achieved a 70.0% accuracy. Adding a machine-readable dictionary and bigram information improved performance to 74.8%, (i.e., just between the "most frequent tag overall" and "most likely tag" baselines obtained on the 200k word corpus).

The present corpus has been the topic of three independent sets of part-of-speech tagging experiments, each running on a differently cleaned version of the corpus, as described in Section 3. The differences in the cleaning strategies (Section 3.6.), as well as differences in the ways the corpus was split up into folds, might explain why the results of the three tagging experiments are not directly compatible.

Tachbelie (2010) achieved an overall accuracy on the 'ELRC' tagset of 84.4% using the Support Vector Machine-based tagger SVMTool (Giménez and Màrquez, 2004) and 82.9% with TnT, Trigrams and Tags (Brants, 2000), a tagger based on Hidden Markov Models, when training on 95% of the corpus and testing on 5%. For unknown words, she received an accuracy of 73.6% for SVMTool and 48.1% for TnT. The set of taggers was later on (Tachbelie et al., 2011) extended with MBT, a Memory-Based Learning tagger (Daelemans et al., 1996), and a CRF-based tagger developed in the toolkit CRF++ (Lafferty et al., 2001). Again SVMTool performed best on unknown words (75.1%) and overall (86.3%) while CRF++ was best on known words (87.6%). Interestingly (and surprisingly), no gain was achieved by combining taggers. However, adding segmentation and reducing the tagset to 16 word-classes did: the results improved by 7% over the board, including a 93.5% overall accuracy for SVMTool on the reduced (16 class) tagset.

Gambäck et al. (2009) report the average 10-fold cross-validated accuracy obtained from three taggers when trained on 90% of the corpus and evaluated on the remaining 10%. Then TnT performed best on known words (over 90% for all three tagsets, incl. 94% on 'SISAY'), but terribly on the unknown ones (only 52% 'ELRC' and 82% on the others), for 85.6% resp. 92.6% accuracy overall. SVMTool out-performed the other taggers both on the unknown words and overall, reaching 92.8% overall accuracy on the 'SISAY' and 'BASIC', and 88.3% on the more difficult 'ELRC' (with 78.9% on unknown words in 'ELRC' and 88.2–88.7% on the others). The third strategy tested was a Maximum Entropy (MaxEnt) tagger (Ratnaparkhi, 1996) as implemented in McCallum's java machine learning package MALLET (`http://mallet.cs.umass.edu`). The MaxEnt tagger performed more in the middle of the road: 87.9% overall accuracy on 'ELRC' and 92.6% on both the smaller sets.

However, the MaxEnt tagger clearly out-performed the other taggers on all tagsets when allowed to create its own, stratified folds: 94.5–94.6% on the'SISAY' and 'BASIC' tagsets, and 90.8% on 'ELRC'. The dramatic increase in the MaxEnt tagger's performance on the stratified folds is surprising, but a clear indication of why it is so important to create a "standard" set of folds, as discussed in Section 4.2.

Gebre (2010) tested both TnT and a CRF-based tagger on the corpus, as well as a version of Brill's transformation-based tagger (Brill, 1995), all on the 'ELRC' tagset. Then TnT and the Brill tagger performed on par, with a 10-fold cross-validated average accuracy of 87.1% and 87.4% respectively, while the tagger based on Conditional Random Fields did clearly better, reaching a 91.0% accuracy, the best result reported for Amharic part-of-speech tagging to date.

6. Conclusions

The paper has described how a 200,863 word part-of-speech tagged corpus of Amharic news texts was created by cleaning, normalising and verifying a publically available manually tagged corpus. The corpus has been marked up with three different tagsets (of 30, 11 and 10 tags each), and also been split into standardized folds for evaluation purposes.

The corpus has been used to train state-of-the-art part-of-speech taggers on Amharic. The best reported results so far show tagging accuracy of around 90% on the most difficult tagset, which is not very encouraging, and not useful for the task of tagging the remainder of the corpus. Rather, figures above 96% would be needed, a level which the best Amharic taggers currently fail to reach even on the easier tagsets.

The efforts on applying machine learning approaches to the task of tagging the corpus has still been useful, though, since it has meant that several errors and tagging inconsistencies in the corpus were spotted and subsequently corrected.

Acknowledgments

This work was carried out while the author was a researcher at SICS, Swedish Institute of Computer Science AB, and has been partially funded by Sida, the Swedish International Development Cooperation Agency, through the Swedish Programme for ICT in Developing Regions (SPIDER).

Thanks to Atelach Alemu Argaw, Lars Asker, Girma Awgichew Demeke, Binyam Gebrekidan Gebre, Mesfin Getachew, Fredrik Olsson, and the staff at ELRC.

7. References

ACL. 2005. *43rd Annual Meeting of the Assoc. for Computational Linguistics*, Ann Arbor, Michigan, June.

Sisay Fissaha Adafre. 2005. Part of speech tagging for Amharic using conditional random fields. In ACL (2005), pp. 47–54. Workshop on Computational Approaches to Semitic Languages.

Atelach Alemu Argaw and Lars Asker. 2005. Web mining for an Amharic-English bilingual corpus. In *1st Int. Conf. on Web Information Systems and Technologies*, pp. 239–246, Deauville Beach, Florida, May.

Roy Bar-Haim, Khalil Sima'an, and Yoad Winter. 2008. Part-of-speech tagging of modern Hebrew text. *Natural Language Engineering*, 14:223–251.

Thorsten Brants. 2000. TnT — a statistical part-of-speech tagger. In *6th Conf. on Applied Natural Language Processing*, pp. 224–231, Seattle, Washington, April. ACL.

Eric Brill. 1995. Transformation-based error-driven learning and Natural Language Processing: A case study in part of speech tagging. *Computational Linguistics*, 21:543–565.

CIA. 2012. *The World Factbook: Ethiopia*. The Central Intelligence Agency, Washington, DC, March. Webpage. https://www.cia.gov/library/publications/the-world-factbook/geos/et.html.

CSA. 2010. *Population and Housing Census of 2007*. Ethiopia Central Statistical Agency, Addis Ababa, Ethiopia, July. Online CD. http://www.csa.gov.et/index.php?Itemid=590.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In *4th Workshop on Very Large Corpora*, pp. 14–27, Copenhagen, Denmark.

Girma Awgichew Demeke and Mesfin Getachew. 2006. Manual annotation of Amharic news items with part-of-speech tags and its challenges. *ELRC Working Papers*, 2:1–17, March.

EACL. 2009. *12th Conf. of the Europ. Chap. of the Assoc. for Computational Linguistics*, Athens, Greece, March.

Björn Gambäck, Fredrik Olsson, Atelach Alemu Argaw, and Lars Asker. 2009. Methods for Amharic part-of-speech tagging. In EACL (2009), pp. 104–111. 1st Workshop on Language Technologies for African Languages.

Michael Gasser, 2009. *HornMorpho 1.0 User's Guide*. Bloomington, Indiana, December.

Michael Gasser. 2011. HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In HLTD (2011), pp. 94–99.

Binyam Gebrekidan Gebre. 2010. Part of speech tagging for Amharic. MSc Thesis, Law, Social Sciences and Communications, Univ. of Wolverhampton, England, June.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *4th Int. Conf. on Language Resources and Evaluation*, pp. 168–176, Lisbon, Portugal, May. ELRA.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In ACL (2005), pp. 573–580.

HLTD. 2011. *Conference on Human Language Technology for Development*, Alexandria, Egypt, May.

Grover Hudson. 1999. Linguistic analysis of the 1994 Ethiopian census. *Northeast African Studies*, 6:89–107.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th Int. Conf. on Machine Learning*, pp. 282–289, Williamstown, Maryland, USA, June.

Saib Mansour. 2008. Combining character and morpheme based models for part-of-speech tagging of Semitic languages. MSc Thesis, Computing Science Dept., Technion, Haifa, Israel.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Eric Brill and Ken Church, editors, *1st Conf. on Empirical Methods in Natural Language Processing*, pp. 133–142, Univ. of Pennsylvania, Philadelphia, Pennsylvania, May. ACL.

Drahomíra Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In EACL (2009), pp. 763–771.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2011. Part-of-speech tagging for under-resourced and morphologically rich languages — the case of Amharic. In HLTD (2011), pp. 50–55.

Martha Yifiru Tachbelie. 2010. Morphology-based language modeling for Amharic. PhD Thesis, Dept. of Informatics, Univ. of Hamburg, Germany, August.

Daniel Yacob. 1997. System for Ethiopic Representation in ASCII (SERA). Webpage. http://www.abyssiniacybergateway.net/fidel/sera-97.html.