

# An Update of the Manually Annotated Amharic Corpus

Pavel Rychlý<sup>1</sup> and Gezahegn Tsegaye Lemma<sup>2</sup>

<sup>1</sup> Faculty of Informatics  
Masaryk University  
Brno, Czech Republic

<sup>2</sup> University of Calabria  
Cosenza, Rende, Italy  
pary@fi.muni.cz

**Abstract.** The paper describes a an update of the manually annotated Amharic corpus WIC 2.0. It lists the problems of the previous version of the corpus and shows that even small changes in the corpus annotation could lead to a higher quality of trained part-of-speech taggers.

**Key words:** text corpus; Amharic corpus; part-of-speech tagging

## 1 Introduction

Amharic language has tens of million native speakers but, in the corpus linguistics, it is one of under-resourced languages. There are not many corpora and language tools for Amharic, but there is a steady progress in creating both language data and tools.

One of very basic tools for natural language processing is a part of speech (PoS) tagger. PoS taggers assign a PoS tag for each word from an input. They usually learn a language model (or a set of rules) using manually annotated corpora. It is very hard to build a tagger without at least small annotated corpus. One of such approaches is described in [5]. Most taggers requires hundreds thousand of tokens for a reliable processing. Building a corpus of that size is expensive in the amount of human work.

In this respect, Amharic language is in a good position, there is the Walta Info Corpus (WIC). It consists of about 210,000 words in 1,065 documents. Texts were taken from the Web news published by the Walta Information Center ([www.waltainfo.com](http://www.waltainfo.com)) in 2001.

There were several attempts to use the WIC Corpus for training automatic part-of-speech taggers, for example [1,8,2]. All of them found that the corpus has many annotation inconsistencies: missing tags, misspelling of tags, multiword expressions and others. Somewhat cleaned version was described in [6] and it was published in the Clarin repository [4].

**Table 1.** PoS tag frequency of the two most ambiguous words in the WIC Corpus

<i>ያህል (about)</i>		<i>በተለይም (specifically)</i>	
ADJ	5	ADJ	5
ADJC	1	ADV	15
ADJP	3	N	11
ADV	141	NC	3
CONJ	3	NP	1
N	1	PREP	2
NC	4	PRON	1
NP	24	UNC	37
NPC	4	V	48
UNC	25	VP	2
VPC	1	VREL	2

## 2 WIC tag set

Amharic language has a rich morphology: Nouns and adjectives are inflected and there are complex rules for deriving verbs. Several part-of-speech tag systems were proposed earlier, all working with about 10 tags for basic part of speech. In some cases, nouns, pronouns, adjectives, verbs and numerals have variants of words with attached prepositions and/or conjunctions. For example, nouns (tag N) could be combined with a preposition as prefix (tag NP), with a conjunction as suffix (tag NC), or with a preposition as prefix and a conjunction as suffix (tag NPC). In total, there are 30 different PoS tags in the WIC Corpus.

### 2.1 PoS tag ambiguity

The WIC Corpus contains very high ambiguity on word form level. There are almost 34,000 types (different word forms including numbers and punctuation marks), almost 20,000 out of them are hapax legomena (occurring only ones). There are more than 4600 types with at least two different PoS tag. The most ambiguous words are *ያህል (about)* and *በተለይም (specifically)* both with 11 different PoS tags. The list of all tags with respective counts in the corpus are listed in Table 1.

Another examples of an ambiguous type is number *10*. In the original version of the corpus, it has 5 alternative PoS tags (each with only one occurrence) in addition to the correct one *NUMCR* (cardinal number). We can guess that many of these PoS tags are plain errors in the annotation. They are not surprising if we consider the annotation process during the corpus building. Annotators wrote the tags on a paper and they were later transcribed into the electronic form.

### 3 Error correction

During our work with the corpus, we have identified ambiguous words and tried to verify PoS tags for them. We have listed 68 words resulting in more than 200 combinations of word-tag. These combinations covers 5000 tokens, we have checked substantial part of all occurrences for each word-tag combination.

We have identified 139 word-tag combination where all hits in the corpus are errors. All the errors were corrected, there are more than 2,300 tokens affected.

In the majority of studied words, most of the occurrences of the given word are correct. For example, word ብለት (*two*) has 150 hits in the corpus, 139 correct and 11 incorrect (with 3 different PoS tags). On the other hand, word ለመስራት (*to work, to make*) has 40 hits, 34 incorrect (PoS tag *NP*) and 6 correct (PoS tag *VP*).

We estimate that there could be about 10 % of errors in the PoS tag annotation.

### 4 Evaluation

The new version of the corpus is different in only a bit more than 1 % of tokens. One can think that this number is too small to have any effect on the PoS taggers trained on the corpus. The more technical changes in our first release of the corpus have very limited effect, and it changed much more tokens. On the other hand the changes described in this paper affect several high frequent words.

To measure the effect of the changes, we have done the same evaluation process as during the previous release of the corpus. We have trained two different PoS taggers and evaluated the accuracy using 10-fold cross validation. We have divided the corpus into 10 parts each containing 20,000 tokens. For each part, we trained a tagger on nine remaining parts, ran the tagger on that part, and compared the result with the manual annotation. The whole evaluation task was done on the Fidel part of the corpus (the Ethiopian script). The evaluation was done before and after the proposed changes.

#### 4.1 TreeTagger

TreeTagger [7] works well for tag-sets with a small number of tags. Both training and tagging is quite fast. The results are listed in Table 2.

We can see that the average accuracy is higher 0.5 percentage points. That is small but significant.

#### 4.2 APtagger

APtagger [3] is a fast and accurate part-of-speech tagger based on the Averaged Perceptron. As neural motivation suggest, the tagger uses several random passes through the training data to learn the model. Each run is a bit different with different model parameters. We have used 10 iterations in training and the resulting accuracy differs in only fraction of percentage point between runs.

The respective results of APtagger are listed in Table 3.

**Table 2.** Accuracy of TreeTagger on ten parts of the WIC Corpus

Part	befor changes	after changes
1	85.1	85.8
2	85.4	86.0
3	85.7	86.3
4	88.2	88.5
5	89.2	89.8
6	86.8	87.3
7	89.9	90.4
8	91.6	91.7
9	89.8	90.3
10	82.3	83.3
<b>Average</b>	87.4	87.9

**Table 3.** Accuracy of APtagger on ten parts of the WIC Corpus

Part	befor changes	after changes
1	80.3	80.9
2	80.7	81.7
3	82.0	82.1
4	83.6	84.2
5	84.4	85.0
6	82.8	83.5
7	85.1	85.6
8	86.5	87.0
9	84.7	85.5
10	79.1	80.0
<b>Average</b>	82.9	83.6

We can see that TreeTagger is significantly better than APtagger. The average accuracy of APtagger is higher 0.7 percentage points after correction of errors in PoS tags.

## 5 Conclusion

In this paper, we have presented the a new release of the WIC Corpus with corrections of PoS annotation of 68 words. The changes affect about 2.500 tokens but even such small number of changes has a significant positive effect on the accuracy of two different PoS taggers.

The new version of the corpus is going to be available again in the Clarin repository.

**Acknowledgments.** This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and by the Grant Agency of CR within the project 18-23891S.

## References

1. Gambäck, B., Olsson, F., Argaw, A.A., Asker, L.: Methods for amharic part-of-speech tagging. In: Proceedings of the First Workshop on Language Technologies for African Languages. pp. 104–111. Association for Computational Linguistics (2009)
2. Gebre, B.G.: Part of speech tagging for Amharic. Ph.D. thesis, University of Wolverhampton Wolverhampton (2010)
3. Honnibal, M.: Aptagger (2013), <https://explosion.ai/blog/part-of-speech-pos-tagger-in-python>, a Good Part-of-Speech Tagger in about 200 Lines of Python
4. Rychlý, P.: Amharic WIC corpus (2016), <http://hdl.handle.net/11234/1-2593>, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
5. Rychlý, P.: Kerneltagger—a pos tagger for very small amount of training data. RASLAN 2017 Recent Advances in Slavonic Natural Language Processing p. 107 (2017)
6. Rychlý, P., Suchomel, V.: Annotated amharic corpora. In: International Conference on Text, Speech, and Dialogue. pp. 295–302. Springer (2016)
7. Schmid, H.: Treetagger | a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart **43**, 28 (1995)
8. Tachbelie, M.Y., Menzel, W.: Morpheme-based language modeling for inflectional language—amharic. Amsterdam and Philadelphia: John Benjamin’s Publishing (2009)