Annotated Amharic Corpora

Pavel Rychlý and Vít Suchomel(✉)

NLP Centre, Faculty of Informatics, Masaryk University,
Botanická 68a, 602 00 Brno, Czech Republic
{pary,xsuchom2}@fi.muni.cz

Abstract. Amharic is one of under-resourced languages. The paper presents two text corpora. The first one is a substantially cleaned version of existing morphologically annotated WIC Corpus (210,000 words). The second one is the largest Amharic text corpus (17 million words). It was created from Web pages automatically crawled in 2013, 2015 and 2016. It is part-of-speech annotated by a tagger trained and evaluated on the WIC Corpus.

1 Introduction

Annotated corpora are quite common even for under-resourced languages but there are languages with tens of million native speakers without high quality text corpora. Amharic is such a case.

Amharic is one of the official working languages of Ethiopia. It is the second most spoken Semitic language in the world with over 20 million native speakers. With so many speakers and being an official language it is hard to believe it counts as an under-resourced language. However, there are not many language resources for Amharic and most of those available are of poor quality, small sized and/or not easily accessible. That is also the case of text corpora. There are several text corpora available (see Sect. 2) but there is only one morphologically annotated corpus of small size and poor quality. One of the reasons for that situation is the special script used for writing Amharic: Ge'ez.

Ge'ez script, also called Fidel in Amharic, is a syllabic script. There are more than 300 characters, each representing a consonant – vowel pair. There are 26 consonant letters combined with 7 or more vowels. The Ge'ez script has also its own symbols for numbers and punctuation. See Table 1 for an example of the Ge'ez characters and Fig. 1 for an Amharic text written in Ge'ez. The Ge'ez script is used also for writing Tigrinya and several smaller languages of Ethiopia. Not all characters are used in all languages –there are characters used only in one language. Ge'ez script is supported by Unicode standard from version 3.0 (1999). Several rarely used characters were added into versions 4.1 (2005) and 6.0 (2010).

Because of the bad support for displaying and writing Ge'ez script on computers there were many attempts to use a transliteration of the script in other alphabets. There are at least ten different transliteration systems in Latin script. Not all of them define mapping of all Ge'ez characters but all are based on a

P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 295–302, 2016.
DOI: 10.1007/978-3-319-45510-5_34

phonetic transcription, hence the differences are not big. The most complete and most different from others is SERA [2]. No accents are required, only ASCII characters (English alphabet) are used. Therefore, it is easy to type SERA on any keyboard. Transliteration of several Ge'ez characters is listed in Table 1. We are using SERA in all our corpora together with the original Ge'ez script.

Table 1. Transliteration of selected Ge'ez characters in SERA system.

ሰ=se	ሱ=su	ሲ=si	ሳ=sa	ሴ=sE	ስ=s	ሶ=so
ሸ=xe	ሹ=xu	ሺ=xi	ሻ=xa	ሼ=xE	ሽ=x	ሾ=xo
ኘ=Ne	ኙ=Nu	ኚ=Ni	ኛ=Na	ኜ=NE	ኝ=N	ኞ=No

2 Existing Corpora

2.1 WIC News Amharic Corpus

Amharic text corpora range from morphologically annotated to parallel corpora. Compared to similar corpora in other (even smaller) languages, all Amharic corpora are small. WIC Corpus [1] is the only manually morphologically annotated corpus. It consists of about 210,000 words in 1,065 documents. Texts were taken from the Web news published by the Walta Information Center (http://www.waltainfo.com) in 2001. A sample of the corpus in displayed in Fig. 1.

<document>
<filename>mes01a1.htm</filename>
<title>ልዩ <ADJ> ዞኑ <N> ያስገነባቸው <VREL> ፕሮጀክቶች <N> 50ሺ <NUMCR> ነዋሪዎችን <N> ተጠቃሚ <ADJ> አደረጉ <V> ። <PUNC></title>
<dateline place="አዳማ" month="መስከረም" date="1/1994/(WIC)/" />
<body>
በአዳማ <NP> ልዩ <ADJ> ዞን <N> በተጠናቀቀው <VP> የበጀት <NP> አመት <N> በ6 ነጥብ 8ሚሊየን <NUMP> ብር <N> ከተጀመሩት <VP> 12 <NUMCR> ፕሮጀክቶች <N> መካከል <PREP> አብዛኞቹ <ADJ> ተጠናቅቀው <V> 50ሺ <NUMCR> ነዋሪዎችን <N> ተጠቃሚ <N> ማድረጋቸው <VN> ተገለጸ <V> ። <PUNC> በልዩ <ADJP> ዞኑ <N> የቴክኒክ <NP> አገልግሎትና <NC> የከተማ <NP> ቦታ <N> አስተዳደር <N> ክፍል <N> ሃላፊ <N> አቶ <ADJ> ግዛው <N> ድንቁ <N> ዛሬ <ADV> ለዋልታ <NP> እንፎርሜሽን <N> ማእከል <N> እንደገለጹት <VP> በልዩ <ADJP> ዞኑ <N> 13 <NUMCR> ቀበሌዎች <N> ተገንብተው <V> ለአገልግሎት <NP> የበቁት <VREL> የመንገድ <NP> ፣ <PUNC> የግብርናና <NPC> የመሰረታዊ <ADJP> ልማት <N> ፕሮጀክቶች <N> ናቸው <AUX> ። <PUNC>

Fig. 1. Example of annotated WIC Corpus

2.2 Morphological Annotation

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. No existing tag-set includes any tags for annotating gender, number and other grammatical categories. In some cases, nouns, pronouns, adjectives, verbs and numerals have variants of words with attached prepositions and/or conjunctions. For example, there are N = noun, NP = noun with a preposition as a prefix, NC = noun with a conjunction as a suffix, NPC = noun with a preposition as a prefix and a conjunction as a suffix. In total, there are 30 different PoS tags in the WIC Corpus.

3 New Amharic Corpora

We have created two new corpora. The first one is a cleaned version of the WIC Corpus, the second one is a new big corpus from the Web. Both corpora are available for querying on the web page of the HaBiT project at https://habit-project.eu/corpora.

3.1 Cleaned WIC Corpus

There were several attempts to use the WIC Corpus for training automatic part-of-speech taggers, for example [3,4,11]. All of them found that the corpus has many annotation inconsistencies: missing tags, misspelling of tags, multiword expressions and others. There were two separate versions of the corpus: one for original Ge'ez script and one with SERA transliteration. In several research papers, they report different number of tokens for each version. We have unified both versions and corrected non matching words either in Ge'ez or SERA depending on a native speaker decision. We have applied all cleaning procedures described in the above mentioned papers.

We have added more unifications of numbers and dates. For example, most of numbers containing decimal point were written as "6 ነጥብ 8" where "ነጥብ" means "point". It is the result of original transcription from hand-written "paper" annotation into computer. Sometimes such string formed one token while there were three tokens in other cases. We have normalised all such occurrences into the correct form (6.8 in this case) with the respective PoS tag. The size of the cleaned corpus is 200,561 tokens. Each token is represented by a word in Ge'ez, its transliteration in SERA and the respective PoS tag.

The cleaned WIC corpus was used to train a PoS tagger. Because of the small number of tags in the tag-set we chose TreeTagger [9], it works very well in such conditions. To evaluate an accuracy of created tagging model we have divided the corpus into 10 parts each containing 20,000 tokens. For each part, we trained a TreeTagger model on nine remaining parts, ran TreeTagger on that part, and compared the result with the manual annotation. The whole evaluation task was done separately on the Fidel part of the corpus and the SERA part, and

for both on data before and after the final cleaning procedure. The results are summarised in Table 2, the average accuracy is 87.4 %. We can see that the final cleaning has not influenced the results much and the performance of TreeTagger is a bit better on the Fidel script than on the SERA transliteration.

Table 2. Accuracy of TreeTager on ten parts of the WIC corpus

Part	Fidel, before	SERA, before	Fidel, after	SERA, after
1	85.1	85.1	85.1	85.2
2	85.4	85.2	85.4	85.1
3	85.7	85.7	85.7	85.7
4	88.2	88.1	88.2	88.1
5	89.1	89.0	89.2	89.1
6	86.6	86.5	86.8	86.6
7	89.9	89.8	89.9	89.9
8	91.5	91.6	91.6	91.7
9	89.7	89.8	89.8	89.9
10	82.3	82.3	82.3	82.3
Average	**87.36**	**87.30**	**87.41**	**87.35**

3.2 Building an Amharic Web Corpus

We have used the following steps to create a big Web corpus: First, adopting the Corpus factory method [6] bigrams of Amharic words from the Crúbadán database[1] [8] were used to query Bing search engine for documents in Amharic. 354 queries yielded 6,453 URLs. URLs of 3,145 successfully downloaded documents were used as starting points for web crawler SpiderLing [10]. URLs of documents crawled in 2013 using a similar approach[2] were added to the set of starting points.

The following language models were created:

- Character trigram model for language detection.[3] 5.2 MB of text from the WIC Corpus and Amharic Wikipedia was used to train the model.
- Byte trigram model for character encoding detection. The model was trained using web pages obtained by the Corpus factory method.
- The most frequent Amharic words from the WIC Corpus wordlist were used as a resource for boilerplate removal tool jusText [7].

The crawler was set to harvest web domains in the Ethiopian national top level domain `et` and other general TLDs: `com`, `org`, `info`, `net`, `edu`. 3.6 GB of

[1] http://crubadan.org/languages/am, by K. Scannell.
[2] We made an unpublished attempt to crawl the Amharic web in 2013.
[3] http://code.activestate.com/recipes/326576-language-detection-using-character-trigrams/, by D. Bagnall.

http responses was gathered in the process. HTML tags and boilerplate paragraphs were removed from the raw data. 42 % of paragraphs were identified as duplicate or near duplicate and removed using tool onion [7]. 66 MB of deduplicated text obtained by the same process in 2013 was added to the data. Sentence boundaries were marked at positions with Amharic end of sentence characters ። and ፧. The final size of the corpus (containing data from years 2013, 2015 and 2016) is 461 MB or more than 17 million words. Finally, the corpus was tagged by TreeTagger with a model trained on the cleaned version of the WIC Corpus. The corpus is called amWaC 16.[4]

3.3 Corpus Properties

Basic properties of corpus sources are summarised in Tables 3 and 4.[5]

We observe the content of news/politic and religious portals has a significant presence in the corpus sources. Since there are only 138 domains with more than 10 documents represented in the corpus, we admit the result collection would benefit from a greater variety of sources.

The most frequent parts of speech in both corpora are nouns and verbs. For details see Fig. 2.

Table 3. The size of corpus structures.

Document count	33,542
Paragraph count	341,327
Sentence count	1,208,926
Word count	17,320,000
Ge'ez lexicon size	955,628
Sera lexicon size	948,553

Table 4. Document count – the most frequent web domains and domain size distribution.

Top level domains		**Web domains**		**Domain size distribution**	
org	14, 582	*.jw.org	6, 717	At least 1000 documents	7
com	11, 927	*.gov.et	4, 599	At least 500 documents	15
et	5, 090	waltainfo.com	2, 818	At least 100 documents	42
net	1, 084	ginbot7.org	2, 666	At least 50 documents	63
cz	724	eotcmk.org	1, 141	At least 10 documents	149
info	85	ethsat.com	894	At least 1 document	573

[4] Amharic 'Web as Corpus' corpus, year 2016.

[5] TLD cz in Table 4 was set by the host server according to the location of the requesting IP address when downloading the data.

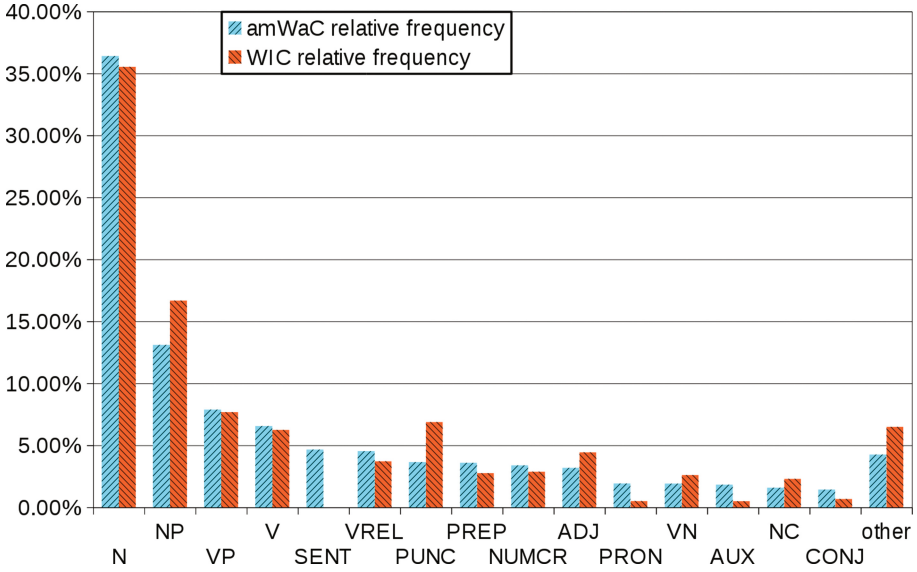

Fig. 2. Relative frequency of tags in both Amharic corpora. (End of sentence token is marked by a PUNCT tag in WIC.)

Table 5. Keyword comparison of amWaC 16 to WIC: words most characteristic for the web corpus, sorted by keyword score.

Word, translation		amWaC 16		WIC		KS
		Count	per million	Count	per million	
ነገር	thing	57927	2855.3	16	79.8	16.4
ነበር	was/where	69074	3404.8	28	139.6	14.6
እንዲህ	like this	27678	1364.3	1	5	13.9
ሥራ	job task	24372	1201.3	0	0	13.0
በ	by/on/at-	32125	1583.5	9	44.9	11.6
ቅዱስ	holy/saint	31392	1547.4	10	49.9	11.0
ዓመት	year	19295	951.1	0	0	10.5
መጽሐፍ	book	17938	884.2	0	0	9.8
እግዚአብሔር	God	17748	874.8	0	0	9.7
ማለት	it means	24235	1194.6	7	34.9	9.6
ቤተ	house-	17902	882.4	1	5	9.4
ሰው	human/man	41861	2063.4	27	134.6	9.2
መንግሥት	government	16300	803.5	0	0	9.0
ክርስቲያን	Christian	16297	803.3	0	0	9.0
ኢየሱስ	Jesus	17968	885.7	2	10	9.0
ዓለም	world	15864	782	0	0	8.8
አበባ	flower	18109	892.6	4	19.9	8.3

Table 6. Keyword comparison of WIC to amWaC 16: words most characteristic for the news corpus, sorted by keyword score.

Word, translation		amWaC 16		WIC		KS
		Count	per million	Count	per million	
ማእከል	centre	455	22.4	1084	5404.8	45.0
እንፎርሜሽን	information	0	0	667	3325.7	34.3
ለዋልታ	to/for Walta	377	18.6	479	2388.3	21.0
ዋልታ	Walta	399	19.7	479	2388.3	20.8
ሚሊየን	million	952	46.9	501	2498.0	17.7
አስታወቀ	he announced	1578	77.8	565	2817.1	16.4
ሃላፊው	the head	177	8.7	315	1570.6	15.4
እንደገለጹት	as they stated	1721	84.8	522	2602.7	14.6
ዘግቧል	he has reported	1688	83.2	470	2343.4	13.3
ተቁመዋል	they pointed out	4	0.2	235	1171.7	12.7
አስታውቀዋል	they announced	1793	88.4	458	2283.6	12.7
ሃላፊ	head	842	41.5	325	1620.5	12.2
ኢንፎርሜሽን	information	754	37.2	292	1455.9	11.3
ወረዳዎች	districts	771	38	274	1366.2	10.6
ገለጻ	briefing	1008	49.7	292	1455.9	10.4
ኢትዮጵያ	Ethiopia	5	0.2	186	927.4	10.2
ምክርቤት	council	0	0	176	877.5	9.8

Tables 5 and 6 show main differences of corpora using keyword comparison: The language is much more formal and the main topic is politics in the news only corpus as expected. Religion related words are noticeable in the WaC corpus. Differences in tokenisation can be observed too, e.g. morpheme ዎ is represented as a separate token in the WaC corpus.

The Keyword Score KS of a word is calculated according to [5] as

$$KS = \frac{fpm_{foc} + n}{fpm_{ref} + n}$$

where fpm_{foc} is the normalised (per million words) count of the word in the focus corpus, fpm_{ref} is the normalised count of the word in the reference corpus and $n = 100$ is the Simple Maths smoothing parameter.[6]

4 Conclusion

We have built a web corpus of Amharic texts comprising of more than 15 million words. To our knowledge it is the largest Amharic corpus for language technology use currently available. We expect the corpus linguistics, lexicography and language teaching in Ethiopia will greatly benefit from such a resource.

We have also cleaned the WIC corpus and unified its Fidel and SERA versions. This resource could be used for building language models (like the TreeTagger model) and for other natural language processing applications for Amharic.

[6] We selected $n = 100$ rather than $n = 1$ to prefer common words over rare words.

A similar approach is being applied to obtain web corpora in other East African languages: Afaan Oromo, Tigrinya and Somali. All corpora compiled within the project are available for browsing and querying by corpus manager Sketch Engine at https://habit-project.eu/corpora. The full source text was not made public because of possible copyright issues.

Acknowledgements. We would like to thank Dr. Derib Ado Jekale from Department of Linguistics, Addis Ababa University for checking seed bigrams of Amharic words, translating key words of the corpus comparison and answering questions about Amharic.

This work has been partly supported by the Grant Agency of CR within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

References

1. Demeke, G.A., Getachew, M.: Manual annotation of amharic news items with part-of-speech tags and its challenges. In: Ethiopian Languages Research Center Working Papers 2, pp. 1–16 (2006)
2. Firdyiwek, Y., Yaqob, D.: The system for Ethiopic representation in ASCII. J. EthioSci. (1997)
3. 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)
4. Gebre, B.G.: Part of speech tagging for Amharic. Ph.D. thesis, University of Wolverhampton, Wolverhampton (2010)
5. Kilgarriff, A.: Getting to know your corpus. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 3–15. Springer, Heidelberg (2012)
6. Kilgarriff, A., Reddy, S., Pomikálek, J., Avinesh, P.: A corpus factory for many languages. In: LREC (2010)
7. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Ph.D. thesis, Masaryk University, Faculty of Informatics (2011)
8. Scannell, K.P.: The crúbadán project: corpus building for under-resourced languages. In: Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, vol. 4, pp. 5–15 (2007)
9. Schmid, H.: Treetagger: a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart **43**, 28 (1995)
10. Suchomel, V., Pomikálek, J., et al.: Efficient web crawling for large text corpora. In: Proceedings of the Seventh Web as Corpus Workshop (WAC7), pp. 39–43 (2012)
11. Tachbelie, M.Y., Menzel, W.: Morpheme-based language modeling for inflectional language–Amharic. John Benjamin's Publishing, Amsterdam and Philadelphia (2009)