Nagaoka Tigrinya Corpus: Design and Development of Part-of-speech Tagged Corpus

Yemane Keleta Tedla Kazuhide Yamamoto Ashuboda Marasinghe
yemane@jnlp.org yamamoto@jnlp.org ashu@kjs.nagaokaut.ac.jp

Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka
Niigata, 940-2188 Japan

1 Introduction

A text corpus is a collection of text data that represents an instance of the language in use. It is a fundamental resource for the development of statistical and corpus-based Natural Language Processing (NLP) tasks. In general, with resource poor languages, the support for working with electronic data is either insufficient or may be not available altogether [11]. We present the development of the first Part-of-Speech (POS) tagged text corpus for a Semitic language, Tigrinya. Tigrinya is spoken by estimated 7 million people in the East African countries of Eritrea and Northern Ethiopia. Other Semitic languages include Arabic, Hebrew, Amharic, Maltese and Tigre. Hebrew and Arabic have a decent amount of resources and research on NLP. However, because of the absence of a Tigrinya corpus, NLP research on the language have been very limited. Therefore, it is imperative to initiate NLP research by developing corpora and thereby language tools for the advancement of information access in Tigrinya.

The raw data used for producing the corpus was collected from a daily newspaper in Eritrea. In order to normalize the corpus, 60 common ways of using clitics were identified from the corpus. In addition, we cleaned the corpus and retained only a few basic punctuation marks. Regarding formatting, the corpus is available in both plain text format and TEI[1] text corpus encoding in XML[2]. A text corpus is more useful when augmented with linguistic information. In light of this fact, part-of-speech information has been added by manually tagging the data. Tagging guidelines used in the process of manual tagging were prepared following a study on Tigrinya grammar. Accordingly, a tagset comprising of 73 tags was designed to encompass three levels of grammatical information: (1) the main POS category, (2) its subcategory or type, and (3) POS affixes (clitics). Finally, these POS labels, were employed in tagging a corpus of 72,080 tokens arranged into 4656 sentences.

2 Tigrinya Language and its morphology

Semitic languages are characterized by rich inflectional and derivational morphology that generates numerous variations of word forms [4]. Tigrinya (also Amharic and Tigre) uses the ancient Ge'ez script as its writing system. The Ge'ez script is one of the few native scripts that are still in active use in the African continent. According to the consonant-vowel syllable, Tigrinya identifies seven vowel sounds which are usually called 'orders'. There are also five-ordered alphabets which are variants of some of the basic 35 consonants. Alto-

[1] TEI – Text Encoding Initiative
[2] XML – Extended Markup Language

gether 275 symbols constitute the Tigrinya alphabet chart known as ‘Fidel’ [9]. Like other Semitic languages, Tigrinya, is also a highly inflected language. Tigrinya verbs are inflected for gender, person, number, case, tense, aspect, mood, etc. Tense-aspect-mood is expressed by affixes that are prefixed, infixed or suffixed to the root word. The verbs have a ‘root-template’ pattern which predominantly consists of triliteral consonants. Subject-verb agreement is enforced by alteration of verb suffixes and/or prefixes. Moreover, negating Tigrinya verbs, nouns or adjectives also require circumfixing the morpheme ኣይ...ን /ayI...nI/[3]. Grammatical clitics that are attached to words further add to the complexity of word morphology. These clitics are mostly proclitics and enclitics of prepositions, conjunctions, possessives and object pronouns [4].

3 The Tigrinya Corpus

The task of developing a Tigrinya corpus is quite challenging for at least two reasons. First, the amount of information available in Tigrinya on the Internet is limited in size and in variation. Second, random crawling of text does not serve the purpose of this research as the text’s meta-data need proper documentation. In the following sub-sections, the process of compiling and annotating NTC is described. The corpus is publicly available[4] on the Internet for research purposes. Processing the corpus is explained in the subsections that follow.

3.1 Data Collection

The raw text was collected from Tigrinya publications of ‘Haddas Ertra’, a newspaper based in Eritrea. The newspaper hosts a range of different topics or domains. News articles published from March 2013 until December 2013 were downloaded from the official site at shabait.com/haddas-ertra. Next, data were randomly selected from different columns of the newspaper as shown in table 1.

Table 1: Distribution of topics

Topic/domain	Articles
Agriculture	10
Business	5
Culture	14
Health	13
History	4
Law	9
Politics	7
Relationship	8
Sport	11
Social	12
General	7
Total	100

It is essential for a corpus to be representative in order to achieve findings that generalize to the language or the specific language domain the corpus represents (Sinclair, 2004). Topic and genre are important parameters that are used to measure how balanced or representative a corpus is [1]. In this regard, although our corpus is a collection of News genre we have attempted to gather data from different topics or domains in order to include more terminologies and structures of the language that could appear in different topics.

3.2 Preprocessing

In the preprocessing phase the raw text was cleaned and formatted into plain text and XML formats. First, unnecessary punctuation marks and foreign scripts were removed. The Corpus was then structured into XML following TEI corpus encoding standards. TEI conformant styles provide an extensive set of markups that support associating meta-information of the entire corpus or an extract of the corpus [6]. Furthermore, in an effort to normalize the corpus, around 60 common methods of cliticizing Tigrinya words were identified from the corpus. This tendency occurs because it is customary to mask laryngeals such as ሕ ‘I’, ሓ ‘a’ or ኢ ‘i’ with an apostrophe. The normalization process takes two forms. In most case, words joined by an apostrophe were separated into full form of their constituent parts. For

[3]Transliteration in this paper uses the uppercase ‘I’ to exclusively mark the epenthetic vowel, traditionally known as ‘sads’

[4]http://eng.jnlp.org/yemane/ntigcorpus

example, ክጽሕፍ’ዩ /kISIHIfI’yu/ ‘he will write’ is resolved into the two words ክጽሕፍ /kISIHIfI/ and እዩ /Iyu/. On the other hand, there are cases were a combination of prepositions such as ኣብ /abI/ ‘in’ , ካብ /kabI/ ‘from’ and pronouns such as እዚ /Izi/ ‘this’, እቲ /Iti/ ‘that’ are found almost fused together without the clitic marker. For instance, considering the word ኣብቲ /abIti/ ‘over there’, while it rarely exists separated as ኣብ እቲ /abI Iti/, there are still some instances of its cliticized form ኣብ’ቲ /abI’ti/ which are merged in the normalization process to ኣብቲ. As such, by minimizing orthographic variations, normalization is expected to simplify the data sparseness problem during training session in various NLP studies such as POS tagging.

3.3 Tagset Design

The experiences of Brown corpus, Penn treebank and related Semitic languages were reviewed during the design of the corpus as well as the tagsets for Tigrinya [2, 3, 7]. Many Tigrinya linguists have classified Tigrinya parts-of-speech into eight major categories [5, 9, 10]. These are verbs, nouns, pronouns, adjectives, adverbs, conjunctions, prepositions and interjections. On the other hand, another study discussed Tigrinya POS by reducing them into five categories [12] . Interestingly, there was a study that classified articles of definiteness as entirely separate POS categories [8].

We followed the most agreed classification, which considers eight major POS tags. The tagset was designed to encompass three levels of POS information. Level-1 represents the tags for the major categories, level-2 is a label for the type of the major tag and level-3 marks if conjunction or preposition is affixed to the main word. In this manner a total of 73 tags were recognized. This design was inspired by previous POS research on Amharic [3]. However, new classes that include level-2 tags for four types of verbs, a proper noun label and a level-1 tag for foreign words were added. Table 2 shows the tags with level-1 and level-2 information. The format of the tagset labels is given as: ‘category_type_preposition and/or conjunction affix’; e.g. the label ‘V_IMF_PC’ represents a Verb (V) of the IMperFective type (IMF) with both a Preposition and a Conjunction (PC) attached to it.

Table 2: Reduced POS tags with category and type information

Category	Type	Label
Noun		N
	Verbal	N_V
	Proper	N_PRP
Pronoun		PRO
Verb		V
	Perfective	V_PRF
	Imperfective	V_IMF
	Imperative	V_IMV
	Gerundive	V_GER
	Auxiliary	V_AUX
	Relative	V_REL
Adjective		ADJ
Adverb		ADV
Preposition		PRE
Conjunction		CON
Interjection		INT
Numeral		NUM
Punctuation		PUN
Foreign Word		FW
Unclassified		UNC

3.4 Manual Tagging

We developed tagging guidelines based on the accounts of three Tigrinya Grammar books [5, 9, 10]. A team of three taggers from Eritrea carried out the manual tagging process using a tagset of 73 tags. The task was conducted in a period of four months, tagging around 72k words. A simple POS annotation tool was developed to aid the manual tagging process. The tool was important in avoiding typographic errors and speeding up the tagging process. In order to ensure tagging consistency, inter-annotator agreement tests were repeated by evaluating the percentage of assignments the taggers agreed on. Subsequently, we refined the tagging guidelines and the taggers were able to agree up to 95% of the time. Out of the 72k words, around 19k (26%) are unique words or types. Furthermore, according to the lexical diversity (token-type ratio), a words gets repeated 3.85 times in the corpus. However, about 6% of the words are

hapaxes. Although there is a large difference, on average a sentence is composed of about 15 words. The distribution of the 12 major tags is shown in figure 1.

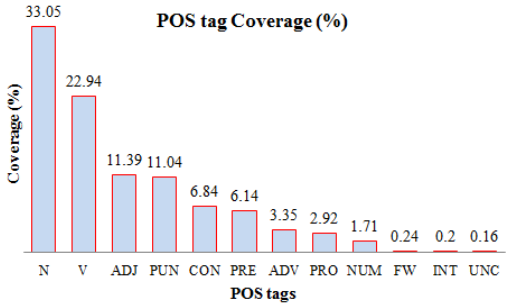

Figure 1: Distribution of POS tags in Nagaoka Tigrinya Corpus

4 Conclusion and Future Works

We presented a study on Tigrinya language conducted with the aim of creating the first part-of-speech tagged corpus for Tigrinya, a low resource language. A corpus of size 72k words was compiled and tagged manually for part-of-speech. The corpus was cleaned, normalized and formatted in TEI corpus encoding. A new tagset of 73 tags was also designed which contains information of the major POS categories, types of the categories and clitics information. In the future, we plan to work on improving both the quality and volume of the corpus. Revising the tagset design and rectifying tagging errors and inconsistencies can improve the quality of the corpus. At present, the corpus contains text from News genre. It is important to collect data from other genres and styles in order to achieve a more representative corpus. The corpus will be an essential resources for many NLP studies.

References

[1] Keh Jiann Chen, Churen Huang, Liping Chang, and HuiLi Hsu. SINICA corpus: Design methodology for balanced corpus. In Language, Information and Computation PACLING 11, 1996.

[2] David Graff and Walker Kevin. Arabic newswire part 1 - Linguistic Data Consertium. https://catalog.ldc.upenn.edu/LDC2001T55, 2001. Accessed: 2014-10-16.

[3] Demeke Girma A. and Getachew Mesfin. Manual annotation of amharic news items with part-of-speech tags and its challenges. Addis Ababa, 2006. ELRC Working Papers, 2:1–17.

[4] Michael Gasser. HornMorpho 2.5 user's guide. Indiana University, Indiana, 2012.

[5] Adi Ghebre. Tigrinya Grammar. Admas Forlag, Stockholm, 2 edition, 2000.

[6] Text Encoding Intiative. TEI P5: Guidelines for electronic text encoding and interchange. http://www.tei-c.org/Vault/P5/current/doc/tei-p5-doc/en/html/, 2013. Accessed: 2013-05-10.

[7] H. Kucera and W. Nelson Francis. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers- Manual of Information. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, 1979.

[8] Sebhatu Gebremichael Kuflu. The basic principles of Tigrinian Language. ForfattaresBokmaskin, Stockholm, 1997.

[9] John Mason. Tigrinya grammar. The Red Sea Press, Inc., New Jersey, 1996.

[10] Amanuel Sahle. A Comprehensive Tigrinya Grammar. The Red Sea Press, Inc., Lawrenceville NJ, 1998.

[11] Oliver Streiter, Kevin Scannell, and Mathias Stuflesser. Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. Springer Science+Business Media, 2007.

[12] Daniel Teklu Reda. Modern grammar of Tigrinya language. Mega Publishing and Distribution PLC, Addis Ababa, 2005.