Universal Dependencies for Amharic

Binyam Ephrem Seyoum[1], Yusuke Miyao[2], Baye Yimam Mekonnen[3]
Addis Ababa University[1,3], National Institute of Informatics[2]
P.O.Box 1176, Addis Ababa[1,3], 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430[2]
binyam.ephrem@aau.edu.et[1], yusuke@nii.ac.jp[2], baye.yimam@aau.edu.et[3]

Abstract

In this paper, we describe the process of creating an Amharic Dependency Treebank, which is the first attempt to introduce Universal Dependencies (UD) into Amharic. Amharic is a morphologically-rich and less-resourced language within the Semitic language family. In Amharic, an orthographic word may be bundled with information other than morphology. There are some clitics attached to major lexical categories with grammatical functions. We first explain the segmentation of clitics, which is problematic to retrieve from the orthographic word due to morpheme co-occurrence restriction, assimilation and ambiguity of the clitics. Then, we describe the annotation processes for POS tagging, morphological information and dependency relations. Based on this, we have created a Treebank of 1,096 sentences.

Keywords: Treebank, Universal Dependencies, Amharic

1. Introduction

In recent years, different language processing applications demand state-of-the-art parsers. Question answering, machine translation, information summarization and similar applications require high-quality parsers. In order to train or develop an efficient parser, it has become an established practice to create a Treebank, linguistically annotated corpus which includes, in most cases, morphological and syntactic annotations. Treebanks play an important role to the research in parsing natural languages. They can also be used in testing linguistic theories and scrutinize corpus-based language analysis. Furthermore, treebanks are essential resources for building and testing data-driven tools such as POS taggers and morphological analyzers where they serve as a gold standard for these tools.

Treebanks have been developed for well-resourced languages in different frameworks such as Phrase Structure, HPSG, and Dependency. However, there are no treebanks for Amharic in any form. In this study, an attempt will be done to create treebanks for Amharic. Apart from developing this resource, the research contributes to the general problem of parsing Morphologically-rich Languages (MRL). In such languages, a dependency relation exists not only between the orthographic words (space-delimited tokens) but also relations within a word itself (Goldberg, Elhadad, and Gurion, 2009). Because of this, clitics attached to orthographic words need to be segmented for proper syntactic analysis. However, automatic segmentation of the prefix and the suffix clitics from the orthographic word in Amharic is problematic due to morpheme co-occurrence restriction, assimilation and ambiguity of the clitics (cf. Section 3 and 4). In this paper, first we discuss clitic segmentation then we describe the creation of the treebanks which are annotated for POS tag, morphological information and dependency relation.

2. Background

Universal Dependencies (UD) project is a collaborative effort to ensure consistent annotations across many languages. This project has benefited from earlier efforts including universal annotation of Google Universal part-of-speech tags (Petrov, Das, and Mcdonald, 2012), morphosyntactic features (Zeman, 2008; Zeman et al., 2012) and Stanford Dependencies (de Marneffe et al., 2014; de Marneffe and Manning, 2008). The objective of UD, as stated in Nivre (2015) is to encourage multilingual parser improvement, cross-lingual learning, and parsing research from a language typology point of view. Even if UD proposes consistent ways of annotations across languages, it does not compromise the unique features of each language. The framework allows language-specific features to be included in annotations. In this paper, we discuss the language-specific features for Amharic.

UD (v2.0) was released on March 01, 2017, with 70 treebanks representing 50 languages (Nivre et al., 2017). All treebanks were annotated with POS tags, morphological features and syntactic relations. Most of them were automatic conversions from one version of treebanks to UD treebanks with manual corrections at some level. The number of sentences were ranging from 600 to 90,000. It also includes some low-resourced languages with a small number of sentences. This demonstrated how low-resourced languages could be benefited from the experience of other languages and contributed to the wider research community. This is also true for Amharic as well. The project encourages more languages to come into the picture.

3. Issues in Amharic Word Segmentation

An orthographic word in Amharic, though, it is delimited by white space, leaves boundaries of lexical or syntactic units unclear. This is because it combines some syntactic words into one compact string of letters. A given orthographic word may attach one or more function words and inflectional morphemes beside the root form. As in Arabic and Hebrew, function words such as prepositions, conjunctions and articles are attached to other content words. This makes an orthographic word in such language function as a phrase, a clause or a sentence. Currently, it has become a trend in Semitic languages to separate function words or clitics as tokens for further linguistic

analysis. For example, አልሰጠሁትም። /ʔalɨsət'tʼəhutɨmm/ “I did not give (it) (to) him.” is written as an orthographic word but it is a full-fledged sentence. This orthographic word encompasses syntactic elements with four parts-of-speech; particle, verb, and two pronominal suffixes. It also expresses three syntactic functions: predicate, subject and direct object.

A syntactic analysis in UD is based on the lexicalist view which says grammatical relations are expressed among syntactic words. It is indicated that practical computational models gain from this approach (de Marneffe et al., 2014). Following this, UD suggests segmentation of function words from content words (Nivre et al., 2016). For example, the above Amharic orthographic word, አልሰጠሁትም። could be segmented into five syntactic forms: አል, ሰጠ, ሁ, ት, ም. However, clitic segmentation is not an easy task in Amharic.

Amharic writing system is said to be ‘syllabic’. Most clitics are vowel forms or at least they begin with a vowel. Since Amharic phonology constrains sequences of two vowels, most clitics undergo phonological changes. The change is also exhibited in the written form where clitics are attached to their host. For proper segmentation, then, we need to recover the hidden form before we segment it. For example, the word ባንድ /band/ "in one”, can be segmented into the preposition በ /bə/ “in” and the numeral አንድ /ʔand/ “one”. However, if we simply segment the first character “ባ” /ba/, the remaining form, ንድ /nd/ will not have meaning.

In addition, the written form in Amharic might lose some grammatical morphemes due to morpheme co-occurrence restriction. For instance, there are some verbs like ተገኘ /təgəǰǰə/ “be found” which can give a sense of passive (which is marked by ተ- /tə-/). When the passive form is used in jussive constructions, ይገኝ /jɨgəǰǰ/ “let it be found”, the passive marker ተ- /tə-/ gets assimilated to the stem initial consonant. Further, the jussive form can serve as input for the imperfective form እሚገኝ /ʔɨmmigəǰǰ/ “that which will be found”. Note that in such imperfective forms, the passive marker ተ- /tə-/ assimilates to the initial consonant of the stem form and the subject marker ይ- /jɨ/ of the jussive form assimilates to the imperfective marker final consonant. The same is true in the case of relative clause, የሚገኝ /jəmmigəǰǰ/ “that which can be found”, where the passive marker ተ- /tə-/ and the subject marker ይ- /jɨ-/ are assimilated and the imperfective marker እም- /ʔɨmm-/ is reduced to -ም /-mm/ only. The process of assimilation and reduction of forms make segmentation of orthographic forms difficult.

Furthermore, some clitic forms can be part of the word without being segmented. In such cases, clitics need context for segmentation; otherwise, they are ambiguous. For example, ከሱ /kəssu/, can mean ‘from him’ or ‘he(honorific)/they lost weight’. It can be segmented into the preposition ከ /kə/ and the pronoun እሱ /ʔɨssu/ for the former meaning but not segmented for the latter meaning.

Segmentation of some clitics may cause other affixes or morphological elements to be separated as well. For instance, we consider the definite marker as a clitic. Unlike Arabic and Hebrew, the definite and the case marker in Amharic are suffixes. When a definite noun appears in an object position, it is marked for the accusative case and the marker follows the definite marker. Thus, segmenting the definite marker has an effect on the status of the case marker that behaves as a clitic. In both Arabic and Hebrew, case markers are treated as morphological features whereas, in Amharic, they are independent syntactic elements. Thus, we have ‘case’ relations rather than morphological features.

When a noun, in Amharic, is modified by an adjective or by other modifiers, the definite marker is attached to one of the modifiers only. In Arabic and Hebrew, such instance is treated as agreement phenomena within the noun phrase. However, in Amharic noun phrase, the definite marker is attached to one of the non-head elements. It could be considered as a phrasal element which can be added to the entire phrase. In our analysis, we treat definiteness at a syntactic level or dependency relation between the noun and the definite marker. The following examples demonstrate our points.

1. መጽሐፉን ሰጠው።
   məs'haf-**u-n** sət't'-ə-w
   book-DEF-ACC give.PRF.-3SGM-3SGM
   “He gave him the book.”
2. ትልቁን መጽሐፍ ሰጠው።
   tɨlk'-**u-n** məs'haf sət't'-ə-w
   big-DEF-ACC book give.PRF.-3SGM-3SGM
   “He gave him the big book.”
3. ጥቁሩን ትልቅ መጽሐፍ ሰጠው።
   t'ɨk'ur-**u-n** tɨlk' məs'haf sət't'-ə-w
   black-DEF-ACC big book give.PRF.-3SGM-3SGM
   “He gave him the big black book.”

In the above examples, the definite marker (-u) and the case marker (-n) are attached to the head noun in (1), but to the adjective in (2) and (3). When the noun phrase expands both markers are attached to the left most element. The noun phrases in (2) and (3) get their definite features from other elements within the phrase. That is why we consider these features as phrasal elements. However, in the segmentation task, since both definite and case markers co-occur, we segment them separately.

Morphemes to be considered as clitics are listed in Binyam, Miyao, and Baye (2016). Following this, we developed a manually segmented data of 2, 300 sentences or 50,520 tokens out of which we selected only 1000 sentences, 12, 039 tokens for the manual annotation of POS tagging, morphological information, and dependency relations.

4. Parts of speech annotation

There have been some works on POS tagging in Amharic (Gamback B., 2012; Martha, Solomon, and Besacier, 2011; Binyam, 2010; Gambäck, Olsson, Argaw, and Asker, 2009; Sisay, 2005). However, the work of Demeke and Getachew (2006), known as the Walta Information Center corpus (WIC), has received much attention among Amharic NLP researchers and has been used for different applications. They propose a 31 tag-set for the manual annotation of a news corpus of 210,000 tokens. The tag-set is based on orthographic words. As a result, they propose a compound tag-set for those words which attach preposi-

tion and/or conjunctions. Since these elements are attached to different lexical categories like nouns, verbs, adjectives, etc, the number of tag-sets has increased. This in return has an effect on the efficiency of automatic taggers trained on the corpus, developed following the proposed tag-set. A recent work by Rychlý and Suchomel (2016) reports an average accuracy of 87.4% of a TreeTagger that is trained and evaluated on WIC.

Besides expected inconsistencies in WIC, which is a manual annotation, such a tag-set has an impact on the performance of an automatic tagger. One impact is, though, they claim to do the task of POS tag, it is beyond the scope of POS tagging. They are trying to give tag-sets for various syntactic constructions, (phrases, clauses and sentences) in addition to a syntactic word. On the other hand, Amharic is a less-resourced and morphologically-rich language where problems of OOV and ambiguities are major bottlenecks. Considering orthographic words for tagging task makes the problems more complex. This is because we are trying to learn several syntactic constructions represented in the orthographic words from a limited corpus.

The other impact is that we miss some information or become confused as the orthography leads to loss of some syntactic information. For instance, in WIC corpus, a separate tag is proposed for relative verbs (VREL). When verbs attach a preposition they are tagged as VP (which means a verb with a preposition). However, when relative verbs attach a preposition, for instance, the relative marker gets deleted due to morpheme co-occurrence restrictions in the language. It is confusing for annotators which tag to use from the orthographic information in such cases. We noted inconsistencies in the tagging of such words in WIC. Some annotators consider the internal structure of a word and tagged them as VREL even if there is a preposition, while others use VP, which contradicts with other similar VP tagged structures. Furthermore, such constructions are also tagged as adjectives (ADJ), considering their modification function in a noun phrase.

In WIC tag sets, it is only the preposition and conjunction that are identified as elements that can be attached to other lexical categories. According to the guideline these elements are attached to nouns, verbs, pronouns, adjective and numerals. However, some adverbs (for instance, ዛሬ /zare/ ‘today’) can attach a preposition and/or conjunction. In addition, the guideline suggests some lexical categories to have sub-classes. Specifically, nouns (verbal noun - VN), verbs (auxiliary - AUX, relative verb - VREL) and numerals (cardinal – NUMCR and ordinal - NUMOR) which have sub-categories with the respective specific tags. However, when these sub-categories attach a preposition or a conjunction, their distinction from the other respective categories cannot be distinguished. This is because the compound tag-sets are used for all categories. For instance, the guideline suggests that a VP tag is used for any verb including auxiliary and relative verbs attaching a preposition. Thus, an auxiliary, other verbs, and relative verbs with a preposition have similar tags as VP. Consequently, an expression tagged as VP following their tag-sets, will have different syntactic structures, i.e. it can be an auxiliary with a preposition or it is a verb or a relative verb with a preposition but tagged similarly. Therefore, the distinction they want to capture by the tags of the sub-categories will not be used when such forms attach a preposition.

The above mentioned problems occur due to the fact that a word is defined as any form that is delimited by a white space. We suggest that for languages like Amharic, clitics should be segmented before tagging and the units for tagging should be syntactic words rather than orthographic words.

When adopting UD, we need to give language- specific information regarding the POS tag-set relevant to Amharic. We need also to provide specific tag-set for some clitics which may as well appear independently. For instance, prepositions and conjunctions can be written separately. For such clitics, we may use the existing tag-sets. However, there are some clitics that need a new tag-set which are result of clitic segmentation.

UD POS	**Amharic tag-set**	**examples**
ADJ	ADJ	ትልቅ “big ”
ADP	ADP	ከ “from”
ADV	ADV	በጣም “very”
AUX	AUX	ኣል “verb to be”
CCONJ	CCONJ	ግን “but”
DET	DET	ይህ “this”
INJ	INJ	ሆ “oh”
NOUN	NOUN	በግ “sheep”
PART	ACC	ን “accusative case”
	NEG	አለ_ሴት “without a woman”
	RLP	የ_መጣ "who came”
	IRLP	እም_ይ_መጣ "who will come”
	NCM	አል_መጣ_ም "He didn’t come”
PRON	PRON	አንተ "you”
	OBJC	ነገር_ኩ_ኣት "I told her”
	SUBJC	ሄድ_ኸ “he went”
	POSM	ቤት_ኢ“my house”
PROPN	PROPN	ካሳ “Kassa”
PUNCT	PUNCT	። "period/fulstop”
SCONJ	SCONJ	ስለ "because”
SYM	SYM	€፣£፣$
VERB	VERB	በላ “eat”
X	X	other

Table1: UD POS tag and Amharic-Specific tag-sets

As can be noted from Table 1, we expand both the particles and the pronouns to handle some clitics that may not have proper tagging after segmentation. Tagging these clitics separately has two advantages. First, segmentation reduces word forms. Due to the morphological structure of the language, word-forms in Amharic are very large. The word-forms even increase with different clitics. Second, it helps to represent syntactic relations between clitics and their host. There is syntactic relation for instance between a preposition and a noun. In the above table, we indicate the mapping between UD tag and Amharic-Specific tag. It is possible to convert Amharic-Specific tags into corresponding UD tags.

5. Morphological annotation

The UD annotation schema defines a set of 21 morphological features across languages. These include Case, Person, Number, Voice and Mood. However, in contrast to the POS tag, the language specification allows treebanks to introduce morphological features that are not included in this universal inventory. This suggests that morphological features can be drawn from the extended compilation of morphological features of other languages (Zeman, 2008).

As we have shown in Section 3 above, due to clitic segmentation some morphological features like the case and the agreement markers are treated as separate forms. Following this decision, case and person features are handled at the syntactic level. Table 2 summarizes the morphological features used in Amharic treebank annotation.

Category	Features	Tag	Description
Nominal	Gender	Mas	Masculine
		Fem	Feminine
		Com	Common gender
	Number	Sing	Singular
		Plur	Plural
		Coll	Collective
Verbal	Verb Form	Conv	Converb
		Inf	Infinite
		Vnoun	Verbal noun
	voice	Pass	Passive
		Mid	Middle
		Rcp	Reciprocal
		Cas	Causative
	Tense	NPas	Future/Present
		Past	Past
	Aspect	Imp	Imperfect
		Perf	Perfect
		Prog	Progressive
		Presp	Prospective
	Polarity	Neg	Negative
		Pos	Affirmative

Table 2: Morphological Features

6. Syntactic annotation

Syntactic dependency types for Amharic are defined in order to be as consistent as possible with the principle of UD. In table 3 below, we provide some samples of typical dependency relations in Amharic. However, the dependency relations for Amharic needs some language-specific information.

One language-specific can be the relation between the subject and/or object clitics and the verb. UD requires the use of the expletive (*expl)* relation for cases of true clitic doubling. In Amharic, the lexical nominal and the clitic may appear in a clause or in a sentence. The nominal will be given the grammatical role of *nsubj*, *obj*, etc., while the clitics will be treated as a pronominal copy of the nominal and will get the role of *expl*. However, when the nominal is dropped, the clitic will get the grammatical roles of *nsubj* or *obj*. Such analysis helps us to handle the case of pro-drop in Amharic. For example, the expression “አለቀሰ" and “እሱ አለቀሰ" are equivalent and can mean “He cried.” The structural difference can be captured using an *expl* relation as indicated in Figure 1.

Relation	Construction in Amharic	→direction
nsubj	ልጁ መጣ።	nsubj(መጣ, ልጁ)
obl	ለልጁ ሰጠው።	obl(ሰጥ, ልጅ)
iobj	ይብተሩኑ ለልጁ ሰጠው።	iobj(ሰጥ,ይብተር)
csubj	የተናግረችው ትርጉም ይሰጣል።	csubj (ይሰጣል, ተናገረችው)
nmod	የእኛ ሀገር	nmod(ሀገር,እኛ)
amod	ትልቁ ልጅ	amod(ልጅ,ትልቅ)
admod	የት መሄድ ትፈልጋለህ ?	advmod (መሄድ, የት)
mark	የመጣው ልጅ	mark(መጣ, የ)
aux	ልጁ ሄዷል።	aux(ሄድ, ኣል)
cop	ልጁ ጎበዝ ነው።	cop(ጎበዝ,ነው)
det	ልጁ (ልጅ_ኡ)	det(ልጅ, ኡ)
acl	የመጣው ልጅ	acl(ልጅ, መጣ)
advcl	ካወቃችሁ ለእስተማሪው ንገሩት።	advcl(እወቅ,ንገር)
ccomp	ጫማውን ልጠግነው እችላለሁ ብሎ ነበር።	ccomp (ብል, ልጠግነው)
expl	ካሳ ለእልማዝ ነገር_ኧ_ኣት	expl(ነገር, ኧ)

Table 3: Some of the dependencies for Amharic

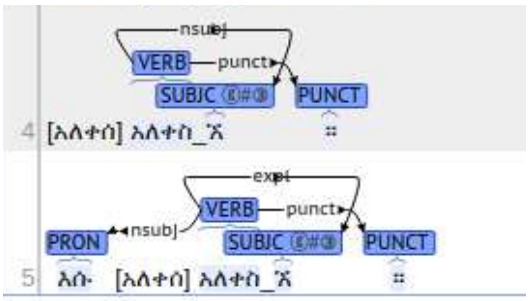

Figure 1: Amharic pro-drop

Another issue is the treatment of converbs. Amharic converbs have features like verbal form, adverbial, non- finite and subordinate. In addition, they modify the verb phrase and uniformly lack specification for most verbal grammatical features like tense, aspect, etc. Thus, we consider them as non-main verbs and the final verb as a main verb. Functionally, converbs may have three functions: serial, consecutive, and co-extensive (Meyer, 2011). They are serial, when they express a chain of actions that constitute one activity and that is concluded by the final verb. They are consecutive when the con-verb expresses an action that takes place earlier than the following verb. They are co-extensive when the action of the con-verb (stative) occurs simultaneously or when they make up one verbal meaning (Desalegn, 2016). We suggest an adverbial modifier (*advmod*) to be used in relation to co-extensive functions. In a structure of subordination, that means both serial and consecutive, we propose to use a sub-relation of

compounds, *compound:svc* (a compound with serial verb construction). Figure 2 and 3 demonstrate this point.

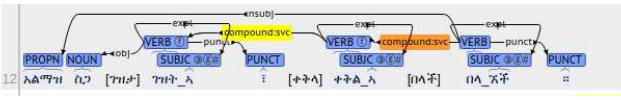

Figure 2: Con-verb with serial construction

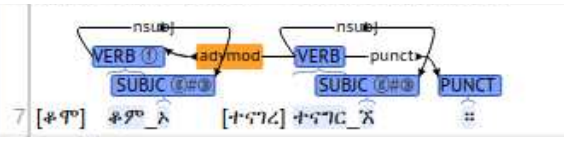

Figure 3: Con-verb with coexistive function

A further language-specific information which is important for Amharic is the treatment of light-verbs in Amharic. Light verb forms do not have a category by themselves. They get their category from their second member. For instance: in construction *sɨbbɨr + allə* – "broken", it is a compound verb as the second member is a verb, where as in *sɨbbɨr + at* - "brokenness", it is a compound noun. The light verb construction is constructed from a light verb and the existential verb. Since, the light verb is semantically null, it cannot be the head. Thus, in such a case, we have decided the copula to be treated as a main verb and the head of the phrase. The relation between the light verb and the main verb is labeled as a compound. Figure 4 shows how light verbs are treated in our UD.

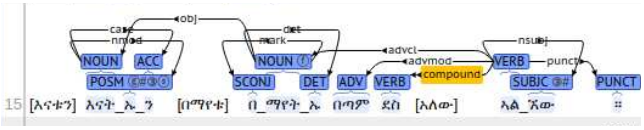

Figure 4: Light-verbs

Copulas in UD are treated as dependent of a lexical predicate (de Marneffe et al., 2014). In Amharic they are used to carry TAM information. However, because we decided to segment subject clitics, the clitics will have syntactic relation with the main verb. The treatment of copula construction is demonstrated in figure 5.

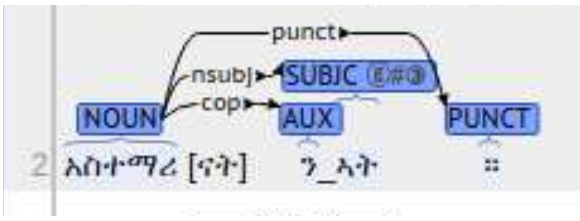

Figure 5: Copula constructions

We have noted that the segmentation of clitics helps us to handle some syntactic relations within a phrase. As we discussed above (cf. Sections 3), there are some clitics that can be attached only once to any one of the constituents within a noun phrase. The phrase gets its features from those clitics attached to the non-head elements. However, in our analysis, the syntactic relations hold between the head and the clitics. The examples (1) to (3) in Sections 3 can be annotated as depicted in figure 6.

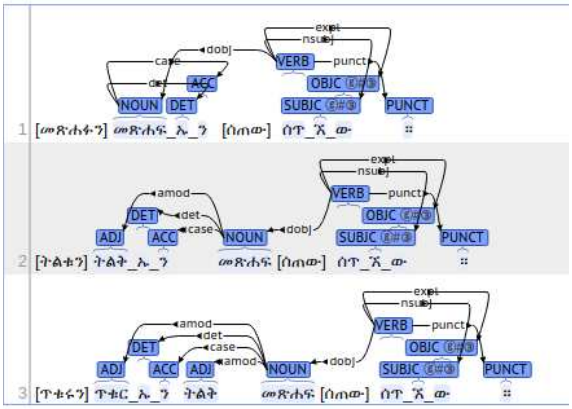

Figure 6: Definite and case marking within NP

7. Corpus and annotation consistencies

In most works of Amharic corpora, data are collected from electronic media, especially from the news media. However, such sources are produced without proper text editing tools like a spell or grammar checker. As a result, errors occur every now and then that require manual editing.

In addition, the Amharic writing system is not standardized. Some people tend to write in phonemic form (the abstract form or what one intends to say) and some tend to write in the phonetic form (what is actually uttered). As a result, there are different forms for a given word in the corpus collected from most electronic media. This makes it difficult to develop a language model. Thus, we focus on collecting sentences from grammar books as they try to cover a variety of grammatical constructions and are consistent with the way the language is written. We have also included other sources like fictions, biographies, religious texts and news.

After sentences were collected from such sources, they were manually corrected for spelling errors. Before the annotation, however, words with clitics were manually segmented. We measured the annotation agreement in the segmentation task. This task has two components. The first task is identifying those words which bear clitics or considered to be complex words. The second task is proper segmentation of the clitics. We measured the agreement using the Kappa measure for both components of the segmentation task. For the first task, we have calculated the number of times that both annotators agree to consider a word as complex, the number of times both agree to exclude, the number of times only annotator one wanted to include a word as complex, and the number of times only annotator two wanted to include a word as complex. Based on this we got the Kappa value of 0.862, which is interpreted as almost perfect agreement.

For the second task in segmentation, we considered those words identified as complex by both annotators. We calculated the number of times that both provide identical segmentation, the number of times that only the first annotator adds more segmentation and the number of times that only the second annotator adds more segmentation. Based on this, we got a Kappa of 0.585

which is considered to be a moderate agreement. Although the agreement was not bad, we asked other linguists to validate their segmentation after the annotators segmented the data, as their input was very important for the further process. Based on the linguists recommendation, we corrected the segmentation data to make it ready for the annotation process.

In the annotation stage, words were annotated for POS, morphological information and syntactic relations which were done manually. Two annotators were trained based on the guideline developed for this purpose. After a series of trainings and updating the guideline, we measured the annotation consistencies for POS tagging and dependency relations using a sample sentence. In order to calculate the Kappa measure for POS agreement, we used a confusion matrix for each tag used in the manual annotation. Accordingly, we got a kappa measure of 0.622, which means there is a substantial agreement between the annotators.

We also measure the annotation consistency for dependency relation. In doing so, we developed a confusion matrix for each dependency relation we used in the annotation. According to the Kappa measure we got, 0.488, there is a moderate agreement between the manual annotators. In order to increase the reliability of the corpus, we have also verified the annotations with two linguists' after the manual annotations were done.

The UD corpus is composed of 1,096 sentences and it contains 8,025 tokens, clitics are not counted as tokens. The data will be released in the upcoming UD version, v2.2[1].

8. Conclusion

We have presented the process of creating Amharic treebanks following the UD annotation scheme. Adopting UD to Amharic needs some kind of decisions regarding the tokens or syntactic words. We have mentioned problems related to clitic segmentation and indicated that Amharic orthographic words may not only bear morphological information but also carry other function elements of syntactic relations. Due to morpheme co-occurrence restrictions, phonological assimilations, and ambiguities, it is difficult to recover syntactic elements from orthographic words. Thus, we suggest that MRL like Amharic segmentation or tokenization of the orthographic word should be the fist step for proper syntactic analysis. For future work, we have a plan to increase the size of segmentation data so that we can develop a machine learning model. In addition, we have a plan to expand the size of the treebank.

Acknowledgements

This project is partially funded by NORHED project called Linguistic Capacity Building-Tools for inclusive development of Ethiopia[2]. We would like thank the University of Oslo for its financial support. We also would like to extend our sincere gratitude to the anonymous reviewers for their valuable feedback which greatly improved this article.

[1]http://universaldependencies.org/

[2]http://www.hf.uio.no/iln/english/research/projects/linguistic-capacity-building-tools-for-the-inclu/

9. Bibliographical References

Binyam, E., Miyao, Y., and Baye, Y. (2016). Morpho-syntactically Annotated Amharic Treebank. In *Proceedings of Corpus Linguistics Fest (CLiF 2016), June 6-10, 2016* (pp. 48–57).

Binyam, G. G. (2010). *Part of Speech Tagging for Amharic*. University of Wolverhampton.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)* (pp. 4585–4592).

de Marneffe, M.-C., and Manning, C. D. (2008). Stanford typed dependencies manual. In *Proceedings of COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Demeke, G. a., and Getachew, M. (2006). *Manual annotation of Amharic news items with part-of-speech tags and its challenges*. *Ethiopian Languages Research Center Working Papers* (Vol. 2).

Desalegn, A. (2016). The Inceptive Construction and Associated Topics in Amharic and Related Languages. Stockholm University.

Gambäck, B., Olsson, F., Argaw, A. A., and Asker, L. (2009). Methods for Amharic part-of-speech tagging. *Proceedings of the First Workshop on Language Technologies for African Languages*.

Gamback B. (2012). Tagging and Verifying an Amharic News Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation.* (pp. 79–84). European Language Resources Association.

Goldberg, Y., Elhadad, M., and Gurion, B. (2009). Hebrew Dependency Parsing: Initial Results. In *Proceedings of the 11th IWPT09* (pp. 129–133). Paris. Retrieved from http://www.aclweb.org/anthology/W09-3819

Martha, Y. T., Solomon, T. A., and Besacier, L. (2011). Part-of-Speech Tagging for Under-Resourced and Morphologically Rich Languages – The Case of Amharic. In *Conference on Human Language Technology for Development* (pp. 50–55). Alexandria, Egypt.

Meyer, R. (2011). The Converb in Amharic. *Journal of Semitic Studies*, (1980), 165–192.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In A. Gelbukh (Ed.), *Proceeding of the 16th International Conference of Computational Linguistics and Intelligent Text Processing, CICLing 2015* (Vol. 9041, pp. 3–16). Springer International Publishing Switzerland.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M., Atutxa, A., … Zhu, H. (2017). Universal Dependencies 2.0. *Http://universaldependencies.org/*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Bullet, M., Ginter, F., Goldberg, Y., … Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1659–1666).

Petrov, S., Das, D., and Mcdonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2012* (pp. 2089–2096).

Rychlý, P., and Suchomel, V. (2016). Annotated Amharic Corpora. In *19th International Conference on Text, Speech and Dialogue (TSD 2016)* (Vol. LNAI 9924, pp. 295–302). Brno, Czech Republic: Springer International Publishing Switzerland.

Sisay, F. A. (2005). Part of Speech tagging for Amharic using Conditional Random Fields. In *Proceedings of the ACL workshop on computational approaches to semitic languages* (pp. 47–54). Association for Computational Linguistics.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)* (pp. 213–218).

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). HamleDT: To Parse or Not to Parse? In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)* (pp. 2735–2741).