(ELRC). It consists of 210,000 tokens collected from 1065 news documents (Demeke and Mesfin, 2006). The corpus is used to develop a stemmer (Argaw, and Asker, 2007), Named Entity recognition (Alemu, 2013), a chanker (Ibrahim and Assabie, 2014). However, the corpus contains some errors and annotation inconsistencies (Gamback, 2012), (Gebrekidan, 2010). Beside the identified problems, they consider orthographic words as their unit of analysis. Function words which are attached to content words are not considered separately. Thus, we cannot use this corpus as it is.

Another important resource is a morphological analyzer called HornMorpho (Gasser, 2011). It is described as "the most complete morphological processing tool for Amharic" (Gamback, 2012). The system can be used to analysis, segment and generate words. The performance was tested on 200 randomly selected words and has been reported to have above 95% accuracy (Gasser, 2011). The tool is developed by taking orthographic words into consideration. As a result, it provides POS, morphological and syntactic information, and other information related to function words that are attached to the word. Even though, the information it provides is important for the analysis of words in isolation, it has to be modified for the purpose of developing a parser. The major focus in the development was on lexical words not on function words. The system gets confused when lexical words attaches more than one function words. For instance, እንደየክልሎች /ʔindəjjəklɨlotʃʃu/ "as to the respective regions", it contains two clitics, ለንደ /ʔɨndə/ and ለየ-/ʔɨjjə/. The system guessed eight analysis whereas when we remove a clitic, it gives the right analysis. However, since clitics are not considered as a separate word, the system does not give any analysis for clitis. Therefore, even though it is a very important tool to check the structure of words, we may not use it for our purpose as it stands.

5. Proposed solutions

In the previous sections we have shown that the existing corpora and tool cannot be used for our purpose due to their limitation of scope or focus. For our purpose, we want to analyze both lexical and function words. Thus, we propose the development of a treebank where both content and function words are separated. In other words, we propose to separate function words or clitics from their phonological host. Even though, the distinction between clitics and affixes are debatable, for our purpose, the following list elements are considered as clitics.

1. Prepositions
2. The Possessive marker or pronominal genitive markers
3. Definite marker
4. Accusative marker
5. Conjunction
6. Negation
7. Auxiliaries
8. Relative pronouns
9. Nominal clause marker
10. Subject and object pronominal agreement markers

The above elements should be separated from content words. To do so, we have collected five thousand sentences from different sources which include grammar books, biographies, news, fictions, science books, law and religions. All the collected sentences were manually checked for spelling errors. These sentences will then be annotated at different levels. Before the annotation, we will decompose words into smaller meaningful units without loss of their basic meaning. As it is indicated above, in Amharic writing system, those listed function words are written together with content words. Thus, we should segment the two. Such segmentation will be done following a guideline which we have prepared.

The guideline gives what should be considered in the manual segmentation. For instance, complex word in the text, that is a combination of a content word and one or more clitics should be embraced by a bracket. This helps to keep track of the input word which is

segmented. When a complex word is segmented, the elements in the orthographic words may not always be the same. They may be modified or reduced in some way. For instance, the word ወደሚገኝ */wədəjəmmigəɲɲ/* "to which that is found" will be segmented into ወደ_የ_እም_ይ_ተገኝ *wədə_jə_ɨmm_jɨ_təgəɲɲ*. From this example we noticed that the orthographic word is a reduced form. When a preposition precedes the complimentizer የ *jə* (relative marker), the form will be reduced into ም *mm*. Thus, we need to keep the input orthographic word using the bracket and show the components that make up the form in the segmentation.

The guidline also provides on how to check wheather a certain form is a clitics or part of the content word. Clitics that we have listed above, in most cases are short forms which may be part of the word. In such cases, they will not be segmented. For instance, the form ከ */kə/* "from" is a preposition. However, it can be part of a content word as in ከበደ */kəbbədə/* "became heavy" or "a personal name". In such cases, the from ከ*/kə/* should not be segemented. This indicates that we cannot apply a certain rule or write a regular expression to automatically segment clitics. Separating clitics, we can say that, requires knoweldge of exisiting words in the laguage.

In addition, the form of the clitics can be changed due to phonological process. This change can also be observed in the orthography. For instance, the prepostion ለ *lə* "to/for" is attached to a content word that begines with the vowel like አሸናፊነት */aʃʃənnafinnət/* "winning", the form of the preposition will be changed.As a result, the form becomes ላሸናፊነት */laʃʃənnafinnət/* " for a winning". If we consider all the variations a clitic may have, it will be problematic to handle all variations. Thus, the guideline suggests to restore to the orginal form in the segmentation. Accordingly, ላሸናፊነት will be segmented into ለ_አሸናፊነት.

The manual segmentation is important to solve some ambiguities observed in the arthography. For instance, some verbs which are relativized can be in active or passive form. This ambiguity occurs because when the relative marker is attached to a passive verb, the passive marker 'ተ' /tə/ will get assimilated to the consonant that begins the word. Thus, we cannot tell whether a relatived verb is an active or passive from the orthography unless we consider the context or the pronouciation. For instance, the word, የሚበላ can read as /jəmmibəlla/ "the one who is eating" as an active form or read as /jəmmibbəlla/ "the one who is being eaten" as a passive form. In the morpholigical anlysis of HornMopho, this is handled by giving both analysis. The following figure shows the analysis of HornMorpho.

```
>>> l3.anal('am', 'የሚበላ')
word: የሚበላ
POS: verb, root: <bl'>, citation: በላ
 subject: 3, sing, masc
 grammar: imperfective, relative
POS: verb, root: <bl'>, citation: ተበላ
 subject: 3, sing, masc
 grammar: imperfective, passive, relative

>>>
```

Figure 1: snap shot taken from HornMorpho analyisis

We notice from figure 1 above that the expression የሚበላ can have two citation forms በላ /bəlla/ "eat" for active and ተበላ /təbəlla/ " being eaten" for passive. The possible interpritation of the expression is given under "grammar" part of the analysis. In our manual anotation, since the segmentation is done for a give sentence which is the context, this expression will be segmented as either as የ_እም_ይ_በል_ኣ or as የ_እም_ይ_ተ_በል_ኣ depending on the contex. The manual segmentation is therefore important to the development of a morphological analyser with a disambugation module for the future.

This stage is the basic and fundamental step where the data is given to three annotators who are linguists and have better understanding of the language for the manual segmentation. Inter-annotators agreement will be checked. After we have reached above 95% inter-annotator agreement, we will assign them a separate data for clitic segmentation. The result of this level will be a corpus of clitics separated from lexical words. It will help us to develop a tokenizer which is a basic tool for the language.

After the segmentation, the corpus will be annotated for POS tag and morphological features. We have compiled 56 POS tag sets based on morphosyntactic properties words. Table 1 summarizes the POS tag sets.

No	tag	Name
1	CN	common
2	ABS	abstract
3	CLN	collective
4	PN	proper
5	VN	verbal noun/infinitive
6	CMN	compound noun
7	PR	personal
8	RF	reflexive
9	DM	demonstrative
10	IN	interrogative
11	IND	indefinite
12	POSP	possessive
13	QAN	indefinite
14	ADJ	other adjectives
15	ADJN	adjective derived from noun
16	ADJV	adjective derived from verbs
17	CADJ	compound Adjective
18	COP	copula
19	VI	intransitive with no complement
20	VIP	intransitive with a complement
21	VT	transitive
22	VTN	transitive with complement
23	VEX	existential
24	VEV	eventive
25	CV	compound verb
26	AUX	auxilary verb
27	ADVT	time
28	ADVP	place
29	ADVD	degree
30	ADVM	manner
31	PREP	adposition
32	POT	postposition
33	CONJ	conjunction
34	SCONJ	subordinating conjunction
35	NUM	cardinal
36	ORD	ordinal
37	INTJ	interjection
38	BG	beginning
39	MD	medial
40	FN	final
41	FW	foreign word, written in other language
42	AC	acronym
43	AB	abbreviation
44	FM	formula
45	EM	emoticon
46	AN	answer
47	NG	negative
48	UC	unclassified
49	SY	symbol
50	MWE	multi word Expression
51	DEF	definite marker
52	ACC	accusative marker
53	GEN	genetive marker
54	NEG	negative marker
55	RLP	relative pronoun
56	POSM	possessive marker

Table 1: POS tag set

The above tags will be revised based on the feedback we will get from the annotators. The list is subjected for modification. In addition, we have listed possible morphological features which words in Amharic can represent. Table 2 lists the morphological features that can be annotated in the treebank.

Basic Categories	Inflection	Type	Tags
Nominal	Gender	masculine	masc
		feminine	fem
		common	com
	Number	singular	sing
		plural	plur
		dual	dual
		collective	coll
	Case	nominative	nom

		accusative	acc
		genitive	gen
	Definite	definite	def
Verb	verb form	infinitive	inf
		gerund	ger
		indicative	ind
		jussive	jus
		question	que
		negative	neg
	Tense	past	past
		present	pres
		future	fut
	Aspect	imperfect	imp
		perfective	perf
		prospective	pro
		progressive	prog
	Voice	active	act
		passive	pass
		reciprocal	rcp
		causative	cau
	Person	first	1
		second	2
		third	3
	Negative	positive /affirmative	pos
		negative	neg
	Agreement	subject	subj
		object	obj
		dative	dat
		applicative	app
	Gender	masculine	masc
		feminine	fem
	Number	singular	sing
		Plural	plu

Table 2: Morphological features

Therefore, the segmented sentences will be annotated for both POS tag and morphological features. This could be done in a semi-automatic way. That means, some of the data like 100 sentences will be manually annotated and then the machine learns the tag and morphological features out of these seed sentences. Then other set of sample sentences will be given to the system to annotate for both type of information. The result will be manually checked and corrected by the annotators. The system again learns from the corrections. In other words, it will be done in iterative ways i.e. manually annotation, training, manual correction, retraining, and annotation (Judge et al., 2006). The result will be used to develop an automatic POS tagger and morphological analyzer.

Finally, we plan to annotate the sentences with grammatical relations using a universal dependency framework (Nivre, 2015). We have identified and compiled potential syntactic relations for Amharic. Table 3 provides potential syntactic relations identified so far.

Category	Relation	Description
Nominal Dependency	adj	adjective
	pred	possessive construction
	app	predicate
	spec	apposition
	cpnd	specification
verbal Dependency	subj	subject of a verb
	pass	passive subject
	obj	object of a verb
	impv	imperative
	pro	prohibition
phrases and clauses	gen	prepositional phrase
	link	PP attachment
	conj	coordinating conjunction
	sub	subordinate clause
	cond	condition
	rslt	result
	conc	concessive
	temp	temporal
	loc	local
	caus	causal
	amd	purpose

Table 3: Dependency relations

Using the above relations, sentences will be semi-automatically annotated for syntactic relations following the same procedures we follow in the above annotations. Consequently, put all the activities together we will have a treebank annotated for all the three information: POS tag, morphological features and dependency relations. Figure (1) is a screen-shot for sample data attempted for a couple of sentences.

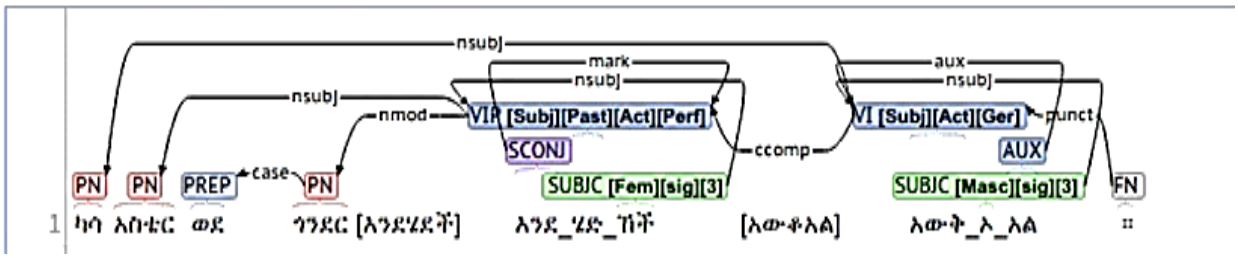

Figure 2: Annotation sample using Brat

In figure 2 we observe a dependency tree for an Amharic sentence. We many notice that complex expressions are embraced by a bracket and their segmentation are indicated following the bracket. When we want to retrieve the orthography we can consider the expression in bracket and when we want to represent the syntactic roles played by the clitics we can consider the segmentation. Furthermore, the morphological features are also indicated together with their tags if a given token has morphological features. We have produced the above kind of representation for some couple of sentences. However, the manual segmentation of the remaining sentences is in progress.

6. Conclusion

We have described an ongoing project that aims at developing a treebank for Amharic. As the language is less resourced and morphologically-rich, we suggest the annotation of the treebank to have three tiers: POS tag, morphological features and syntactic relations. Before the annotation is done, orthographic word needs to be segmented if it has clitics. We suggested that the minimal unit for our analysis should be syntactic words, i.e. both content word and functional words.

Acknowledgments

This project is partially funded by the Director for Research of AAU under *Adaptive Problem-Solving Research grant* and NORHED fund under *Linguistic Capacity Building-Tools for inclusive development of Ethiopia,* (http://www.hf.uio.no/iln/english/research/projects/linguistic-capacity-building-tools-for-the-inclu/). We want to thank both institutions for their support. We also would to extend our appreciations to the three anonymous reviewers for their feedback that greatly improved the article.

References

Abeba Ibrahim and Yaregal Assabie. (2014). Amharic Sentence Parsing Using Base Phrase Chunking. *CICLing 2014*, (pp. 297-306).

Alemu, Besufikad. (2013). A Named Entity Recognition for Amharic . *MA Thesis, Addis Ababa University*.

Amsalu, Saba and Demeke, Girma A. (2006). Non-concatinative Finite-State Morphotactics of Amharic Verbs.

Argaw, Atelach Alemu and Asker, Lars. (2007). An Amharic Stemmer : Reducing Words to their Citation Forms. *Proceedings of the 5th Workshop on Important Unresolved Matters, pages* , (pp. 104 -110). Prague, Czech Republic: Association for Computational Linguistics.

Binyam Gebrekidan. (2010). Part of Speech Tagging for Amharic. *Masters Thesis, University of Wolverhampton*. Wolverhampton.

Carroll, John. (2000). Statistical Parsing. In N. T. R. Dale (Ed.), *Handbook of Natural Language Processing* (pp. 525–544).

Dehadari, J., Tounsi, L. and Genabith, J. (2011). Morphological Features for Parsing Morphologically- rich Languages: A Case of Arabic . *Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically-Rich Language (SPMRL 2011)*, (pp. 12-21). Dublin Irelan.

Dehadari, Jon, Tounsi, Lamia and Genabith, Josef. (2011). Morphological Features for Parsing Morphologically-rich Languages: A Case of Arabic. *Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically-Rich Language (SPMRL 2011)*, (pp. 12-21). Dublin Ireland.

Demeke , Girma Awgichew and Getachew , Mesfin. (2006). Manual annotation of Amharic news items with part-of-speech tags and its challenges. *Ethiopian Languages Research Center Working Papers, 2*, 1–16.

Frank, Aneette; Hinrichs, Erhard;. (2012). Treebanks : Linking Linguistic Theory to Computational Linguistics. *Linguistic Issues in Language Technology*, *7*(1).

Gamback, Björn. (2012). Tagging and Verifying an Amharic News Corpus. *the 8th International Conference on Language Resources and Evaluation (ELRA). Workshop on Language Technology for Normalisation of Less-Resourced Languages.* Istanbul, Turkey.

Gambäck, Björn; Olsson, Fredrik; Atelach, Alemu Argaw; Asker, Lars. (2009). Methods for Amharic part-of-speech tagging. *Proceedings of the First Workshop on Language Technologies for African Languages*, (pp. 104-111).

Gasser, Michael. (2010). A Dependency Grammar for Amharic. *Workshop on Language Resource and Human Language Technologies for Semitic Languages.*

Gasser, Michael. (2011). HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. *Conference on HUman Language Technology for Development.* Alexandria, Egypt.

Habash, Nizar (2010). *Introduction to Arabic Natural Language Processing.* Morgan & Claypool,.

John Judge, Aoife Cahill, Josef van Genabith. (2006). QuestionBank: Creating a Corpus of Parse-Annotated Questions. *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics.* Sydney, Australia: ACL.

Maamouri, Mohamed and Bies, Ann. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 2-9.

Nivre, Joakim. (2006). Two Strategies for Text Parsing. *Journal of Linguistics, 19*, 440–448.

Nivre, Joakim. (2008). Treebanks. In M. a. Kytö, *Corpus Linguistics: An International Handbook* (pp. 225-241). Mouton de Gruyter.

Nivre, Joakim. (2015). Towards a Universal Grammar for Natural Language Processing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 3–16). Switzerland: Springer International Publishing. doi:10.1007/978-3-642-28601-8

Nivre, Joakim; Hall, Johan; Sandra, Kubler; Mcdonald, Ryan; Nilsson, Jens; Riedel, Sebastian & Yuret, Deniz. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, (pp. 915–932). Prague.

Rens Bod ; Remko Scha ; Khalil Sima. (2012). Data Oriented Parsing. (R. Bod , R. Scha , & K. Sima, Eds.) *CSLI Studies in Computation, 14*(4), 472–476.

Tsarfaty , Reut; Seddah, Djame; Kubler , Sandra; Niver, Joakim;. (2013). Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Association for Computational Linguistics*, *39*(1), 15 - 22.

Tsarfaty, Reut. (2013). A Unified Morpho-Syntactic Scheme of Stanford Dependencies. *Proceedings of ACL*.

Tsarfaty, Reut; Seddah , Djame; Goldberg, Yoav; Kubler , Yannick; Candito , Marie; Foster , Jennifer; Versley , Yannick; Rehbein, Ines; Tounsi, Lamia. (2010). Statistical Parsing of Morphologically Rich Languages (SPMRL): What, How and Whither. *The Proceedings of the NAACL HLT 201 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, (pp. 1-12). Los Angeles, California.

Tsarfaty, Reut; Seddah , Djame; Goldberg, Yoav; Kubler , Yannick; Candito , Marie; Foster , Jennifer; Versley , Yannick; Rehbein, Ines; Tounsi, Lamia. (2010). Statistical Parsing of Morphologically Rich Languages (SPMRL): What, How and Whither. *The Proceedings of the NAACL HLT 2010 First Workshop on Statical Parsing of Morphologically-Rich Languages*, (pp. 1-12). Los Angles, California.