Development of an Amharic Dependency Treebank: Annotated Corpus Construction and Linguistic Annotation[1]

Keywords: Amharic, Universal Dependencies, Treebank, Linguistic Annotation

As part of my Master's in Language Technology (MLT) thesis, I propose to undertake the development of an enhanced Universal Dependencies (UD) treebank for Amharic. The proposed project aims to create an enhanced Universal Dependencies (UD) treebank specifically tailored for the Amharic language. Amharic, one of the major languages of Ethiopia, lacks comprehensive linguistic resources, hindering the development of natural language processing tools and applications. By constructing a high-quality UD treebank, this project seeks to fill this gap and facilitate research in syntactic analysis, machine translation, and other NLP tasks for Amharic.

The treebank will include manually annotated syntactic structures following the UD annotation guidelines, encompassing a diverse range of text genres and linguistic phenomena specific to Amharic. Additionally, the treebank will be enriched with morphological and syntactic features to capture the intricacies of the language's grammar and syntax accurately.

Participation in the training school will provide essential training and guidance on best practices for constructing and annotating UD treebanks (de Marneffe et al., 2008). Through workshops, tutorials, and collaborative sessions, I aim to gain expertise in linguistic annotation and syntactic analysis, which will be instrumental in developing a high-quality treebank for Amharic. Furthermore, interaction with experts and fellow participants will offer valuable insights and feedback to refine the annotation process and ensure the quality and consistency of the treebank.

Questions for the brainstorming:

1) What are the challenges specific to annotating syntactic structures in Amharic, and how can they be addressed effectively?
2) How can we leverage existing linguistic resources and tools to expedite the annotation process while maintaining annotation quality?

[1] Dawit J, Master's Student in Language Technology, University of Gothenburg
Email: gusjembda@gu.se (or dawitjtilahun@gmail.com)

3) What strategies can be employed to ensure cross-linguistic consistency and interoperability with other UD treebanks?
4) How can the treebank be extended to cover diverse linguistic phenomena and text genres prevalent in Amharic?

This is a draft timetable for the " Development of an Amharic Dependency Treebank: Annotated Corpus Construction and Linguistic Annotation" project from January 1, 2025, to May 30, 2025:

Activity / Time	January		February		March		April		May	
	W1-2	W3-4	W 1-2	W3-4	W 1-2	W3-4	W1-2	W3-4	W 1-2	W3-4
1. Define project goals, objectives										
2. Set up project management tools										
3. Identify and gather a diverse range of Amharic texts										
4. Ethical considerations										
5. Review existing annotation guidelines										
6. Adapt and customize annotation guidelines										
7. Conduct a pilot annotation										
8. Evaluate the effectiveness of annotation guidelines										
9. Begin manual annotation of the corpus with dependency structures										
10. Maintain documentation of annotation										
11. Consolidate the annotated data										
12. Prepare documentation										
13. Conduct a comprehensive review of the entire project										
14. Prepare final project reports										

Reference

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

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

Demeke, G., & 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*.

Ephrem, B., & G, G. (2010). Part of Speech Tagging for Amharic. *University of Wolverhampton*.

Ephrem, B., Miyao, Y., & Mekonnen, B., Y. (2016). Morphosyntactically Annotated Amharic Treebank. *Proceedings of Corpus Linguistics Fest (CLiF 2016)*, 48–57.

Ranta, A. (2024). *Computational Grammar An Interlingual Perspective*.

Seyoum, B. E., Miyao, Y., & Mekonnen, B. Y. (2018). Universal Dependencies for Amharic. *International Conference on Language Resources and Evaluation*.