

Title: Advancing Universal Dependencies: Towards Enhanced Syntactic Annotation in the Albanian Treebank

Applicant: Enriketa Sogutlu, Beder University College, Albania

Keywords: Universal Dependencies, Albanian Treebank Annotation, Syntactic Analysis, Syntactic annotation

Description of the Resource

The proposed project aims to advance the field of Universal Dependencies (UD) by enhancing the syntactic annotation of the Albanian Treebank within the UD framework. This project builds upon a comprehensive exploration of UD and syntactic annotation methodologies. Through systematic error identification, correction, and peer review, the project seeks to refine syntactic annotation practices and enrich linguistic resources for the Albanian language. In a second phase it would elaborate on the inclusion of sentences containing clitics and combined clitics and their annotation.

The Albanian Treebank represents a valuable linguistic resource for syntactic analysis within the UD framework. It also serves as a cornerstone resource for linguistic research and natural language processing (NLP) applications in the Albanian language. However, like many treebanks, it faces challenges such as inconsistencies in syntactic annotation and classification errors which impact the accuracy and reliability of downstream analyses. The proposed project aims to address these challenges by systematically reviewing and refining the syntactic annotation of the Albanian Treebank. Through a rigorous process of error identification, correction, and peer review, the project seeks to enhance the quality and consistency of syntactic annotations, ensuring alignment with established Universal Dependencies (UD) guidelines. By leveraging insights gained from the exploration of UD principles and methodologies, the project will strive to enrich the Albanian Treebank with accurate and reliable syntactic annotations, thereby enhancing its value as a linguistic resource for both research and practical applications as well as educational purposes.

Explanation of Participation in the Training School

Participation in the training school will play a crucial role in advancing the project by providing essential training and guidance in UD annotation methodologies. The training sessions will offer opportunities to collaborate with experts and fellow participants, exchange ideas, and explore innovative approaches to syntactic analysis. Additionally, the training school will facilitate access to resources and expertise necessary for the successful completion of the project. The knowledge and insights gained during the training sessions will contribute to the project's objective of enhancing the quality and consistency of the syntactic annotation in the Albanian Treebank in the UD.

Open Questions for Brainstorming Hackathon

How can we address the identified issues of insufficient syntax content and categories in the Albanian Treebank, particularly in terms of sentence and clause types?

What strategies can be employed to improve the annotation of the verb "to be" and its classification as auxiliary or main verb?

How can we enhance the classification of different types of determiners, particularly particles preceding adjectives, in the Albanian Treebank?

How can clitics and combined clitics be properly classified and annotated?

What educational and research opportunities exist for leveraging the enhanced syntactic annotation of the Albanian Treebank, particularly in comparative linguistic analysis with other languages in the UD?

Project Phase:

The project has transitioned from preliminary exploration of some of the sentences in the treebank to a more refined investigation of all the sentences with initial results informing the refinement of annotation guidelines and methodologies. The participation in the training school will further advance the project by providing essential training, fostering collaboration, and facilitating access to expertise and resources. Ultimately, the project aims to contribute to the ongoing development of linguistic resources for the Albanian language within the Universal Dependencies framework, with the potential for broader applications in linguistic research and education.

Conclusion

In conclusion, the proposed project represents a new step towards advancing syntactic annotation practices within the Universal Dependencies framework, and in particular the Albanian Treebank. By addressing existing challenges and using expertise gained through collaboration and training, the project aims to enhance the quality and reliability of linguistic resources for the Albanian language. Through continued collaboration and innovation, the project seeks to contribute to the broader goals of UD research and promote cross-linguistic understanding and analysis.

References

- Kabashi, B., & Proisl, T. (2016, May). A proposal for a part-of-speech tagset for the Albanian language. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4305-4310). <https://aclanthology.org/L16-1682.pdf>
- Kabashi, B., Herbst, T., & Götz-Votteler, K. (2007). Pronominal clitics and valency in Albanian: A computational linguistics perspective and modelling within the LAG-framework. *Valency. Theoretical, Descriptive and Cognitive Issues*, 339-352.
- Kabashi, B. (2016). Building an Albanian text corpus for linguistic research. *Kumtesë në konferencën "Corpus-Based Approaches to the Balkan Languages and Dialects*, 5-7.
- Misini, A., Canhasi, E., & Krrabaj, S. (2020, September). Albanian syntactic parsing. <https://repository.ukim.mk/bitstream/20.500.12188/9490/1/albanian-syntactic-parsing.pdf>
- Reppen, R. (2010). Building a corpus. *The Routledge handbook of corpus linguistics*, 31-37. <https://www.toroscevir.info/wp-content/uploads/2022/03/ch3.pdf>
- Ramadani, N. (2020). CONTRASTIVE ANALYSIS OF RELATIVE CLAUSES IN ALBANIAN AND ENGLISH—STRUCTURE AND USAGE. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, 9(11), 31-35. <https://www.anglisticum.org.mk/index.php/IJLLIS/article/view/2138/2559>
- Toska, M., Nivre, J., & Zeman, D. (2020, December). Universal dependencies for Albanian. In Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020) (pp. 178-188). <https://aclanthology.org/2020.udw-1.20.pdf>
- <https://www.besim-kabashi.net/publications.html>
- <https://aclanthology.org/people/b/besim-kabashi>