Development of an Enhanced Universal Dependencies Treebank and Multiword Expression Annotation for Standard Albanian Language

Anila Cepani[1], Adelina Cerpja[2], Nelda Kote[3], Alba Haveriku[3]
[1] University of Tirana
[2] Academy of Sciences of Albania
[3] Polytechnic University of Tirana

The proposed project aims to enhance and enlarge the annotated corpus for the Standard Albanian language presented in (Kote, Cepani, & Haveriku, 2024) and to construct a corpus annotated with Multiword Expressions (MWEs) for the Standard Albanian language.

The corpus proposed in (Kote, Cepani, & Haveriku, 2024), can be considered the biggest treebank syntactically annotated using Universal Dependencies (UD) and the first one by linguistic experts. The treebank has 1.300 annotated sentences and during the annotation we have faced a lot of challenges. There are some annotations decisions to be discuss and find a better solution according to UD and Albanian grammar. The treebank need to be enlarged and new sentences to be annotated.

Multiword expressions play a crucial role in natural language processing tasks such as machine translation, sentiment analysis, and information retrieval. However, existing linguistic resources for Albanian often lack comprehensive annotations for MWEs, hindering the development and performance of language technology applications.

The planed tasks are:

- Data Collection: Gathering diverse texts in Albanian from various sources, including news articles, literary works, and online forums.
- Annotation: Annotate the existing treebank with MWE. Annotate the new selected sentences with syntactic annotation and MWE.
- Quality Assurance: Implementing rigorous quality control measures to ensure the accuracy and consistency of the annotation of the treebank.
- Resource Integration: Incorporating the annotated corpus into existing linguistic repositories or making it publicly available for use by researchers and developers.
- AI model training: The created treebank will be used to train an AI.

How participation in the training school will benefit the project

Participation in the UNIDIVE Summer School offers a unique opportunity to enhance the quality and effectiveness of the proposed project. By attending courses on corpus annotation infrastructure and multiword expressions, the team will gain valuable insights and practical skills relevant to enhance the annotation using UD and to development the MWE annotation for the Standard Albanian language. Interacting with experienced instructors and fellow

participants will enable us to exchange ideas, discuss challenges, and explore best practices in UD and MWE annotation.

Moreover, the brainstorming hackathon provides a platform to solicit feedback, address potential obstacles, and brainstorm innovative solutions to enhance the quality and utility of the annotated corpus.

Open questions for the brainstorming hackathon:

- How can we define and annotate different types of multiword expressions in Albanian, considering its rich morphology and syntax?
- What strategies can be employed to handle the ambiguity and variability of multiword expressions across different contexts and linguistic registers?
- How can we leverage existing resources and tools to expedite the annotation process while maintaining high-quality standards for MWE identification?
- What are the potential applications and downstream tasks that could benefit from the availability of a comprehensive MWE annotated corpus for Albanian?

Project phase statement

Planning (Phase 1): The project is currently in the planning phase, wherein initial preparations, including data collection strategies and annotation guidelines development, are underway. Participation in the UNIDIVE Summer School will inform the project's design and implementation strategies, laying the groundwork for subsequent phases of data collection, annotation, and integration into existing linguistic repositories.

Bibliography

Kote, N., Cepani, A., & Haveriku, A. (2024, February 8-9). Universal Dependencies Treebank for Standard Albanian. *Unidive, 2nd General Meeting University of Naples L'Orientale*. Naples, Italy.