Moldova Training School with UniDive

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

UT

Date: 8-12 July 2024

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

Anila Çepani[1], Adelina Çerpja[2], Nelda Kote[3], Alba Haveriku[3]

[1] University of Tirana, [2]Academy of Sciences of Albania, [3]Polytechnic University of Tirana

Project Overview: Enhance and expand the existing treebank for the Standard Albanian language and add annotations for Multiword Expressions (MWEs).

Achievements:

- **Treebank:** 1,300 annotated sentences.
- **Challenges**: Annotation decisions need to be improved according to UD and Albanian grammar.
- **Expansion Needed**: Annotate new sentences.

Planned Tasks:

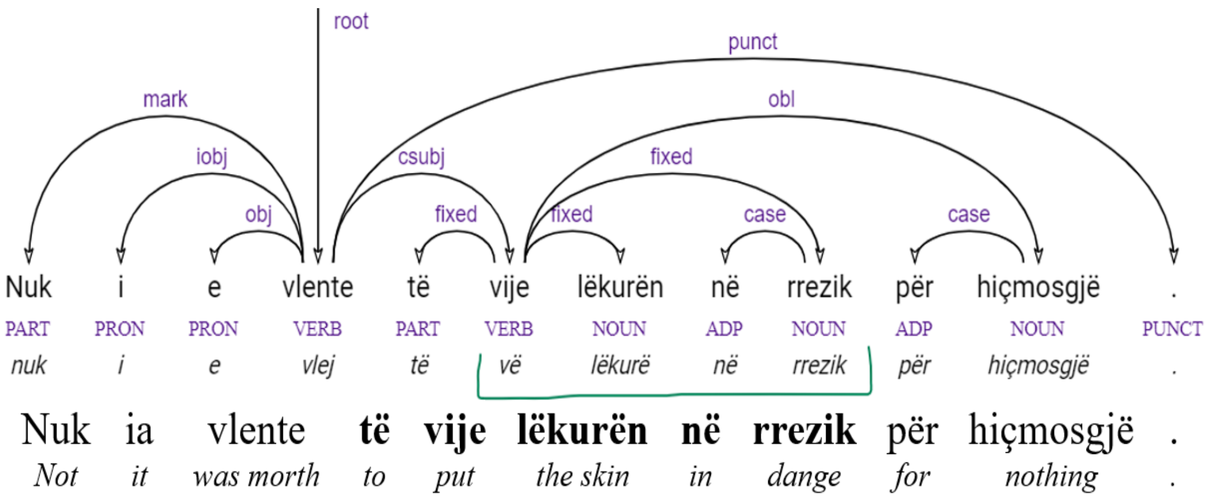
Data Collection: Collect Albanian texts from various sources.

Annotation:

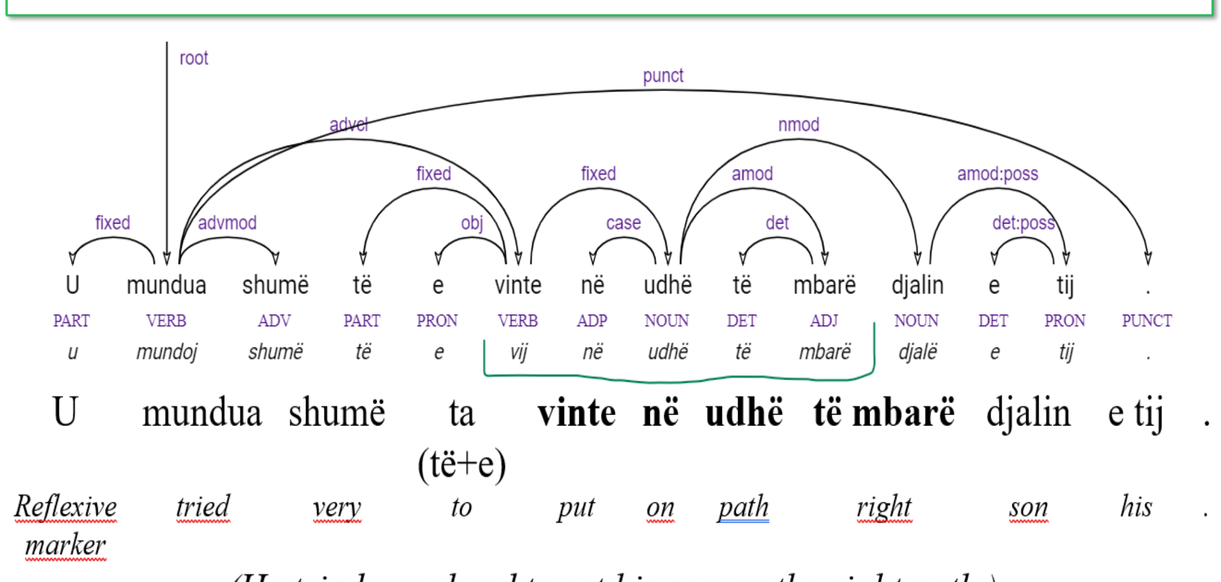
- Annotate MWEs in the existing treebank.
- Annotate MWEs in new sentences.

Quality Assurance: Ensure annotation accuracy and consistency.

Resource Integration: Make the treebank publicly available.



(It wasn't worth risking your skin for anything.)



(He tried very hard to set his son on the right path.)

Training School Benefits:

- **Skills Enhancement**: Gain insights on UD and MWE annotation.
- **Interaction**: Exchange ideas and discuss challenges.
- **Hackathon**: Brainstorm solutions and improve corpus quality.

Hackathon Questions:

1. How to define and annotate MWEs in Albanian, considering its rich morphology and syntax?
2. What are the strategies to handle the ambiguity and variability of MWEs?
3. How can we leverage existing resources for efficient annotation?
4. What applications benefit from a comprehensive MWE-annotated corpus?

Project Phase:Planning (Phase 1)

Initial preparations, including data collection strategies and annotation guidelines, are underway. Participation in the UNIDIVE Summer School will inform the project's design and implementation.

European Union