WG1, WG3

UniDive

2nd General Meeting

University of Naples L'Orientale

Naples, Italy, 8-9 February 2024

https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Universal Dependencies Treebank for Standard Albanian

Nelda Kote[1], Anila Çepani Sema[2], Alba Haveriku[1]

[1]Polytechnic University of Tirana, Tirana, Albania

[2]University of Tirana, Tirana, Albania

We present the Universal Dependencies treebank for the Standard Albanian language, annotated by linguistic experts with the support of information technology professionals. The annotated treebank contains 85,000 tokens (4,000 sentences), of which 25,000 tokens (1,300 sentences) are annotated with syntactic dependencies, part-of-speech tags, morphological features, and lemmas, while the remaining part lacks syntactic dependency annotations. x

Annotation Decisions

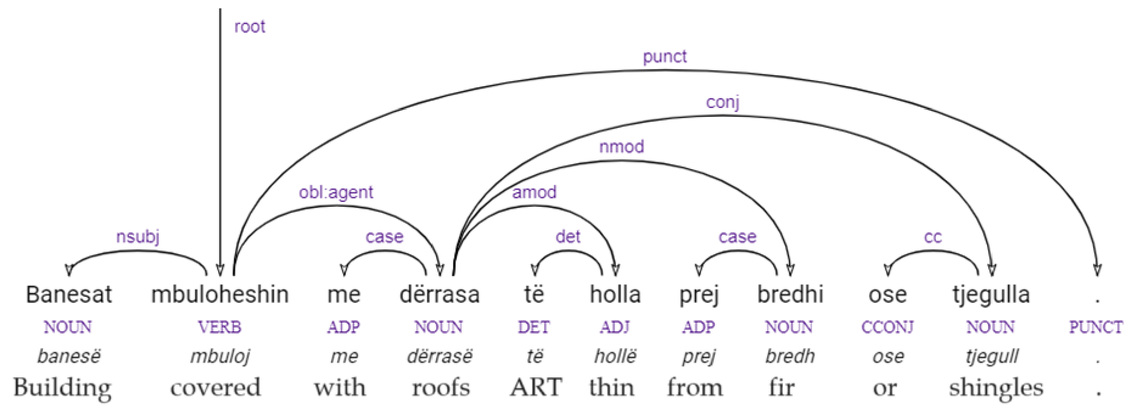
- Sentence segmentation: performed using white space and punctuation marks as boundaries.
- Words segmentation within a sentence: performed using white space and punctuation marks as boundaries poses challenges in labelling analytical grammatical forms and expressions in Albanian, a synthetic-analytical language with both synthetic and analytic features.
- Lemmatization: Linguistic experts determine word lemmas using the Albanian National Dictionary (ASHSH, 1998, 2002, 2006), considering the context and meaning in sentences to prevent ambiguity.
- Part-of-speech tags: A total of 17 part-of-speech tags from the UD tag set are utilized.
- Morphological features: corresponding morphology features based on the word's part-of-speech tag.
- Syntactic annotation: A total of 32 syntactic tags from the UD tag set are utilized.

POS and Morphological feature tags

	POS tag	Morphological features
verb	VERB/AUX	mood, time, person, number, voice; verb form only in case of participle
noun	NOUN	gender, number, case, definiteness
proper noun	PROPN	gender, number, case, definiteness; Abbr in case of abbreviation
adjective	ADJ	gender, number, case, degree
pronoun	PRON	depends on the type (case, number, gender, person, prontype)
adverb	ADV	AdvType
numeral	NUM	NumType
interjection	INTJ	
preposition	ADP	case
particle	PART	
conjunction	CCONJ/ SCONJ	
articles	DET	gender, number, case and prontype
symbols	SYM	

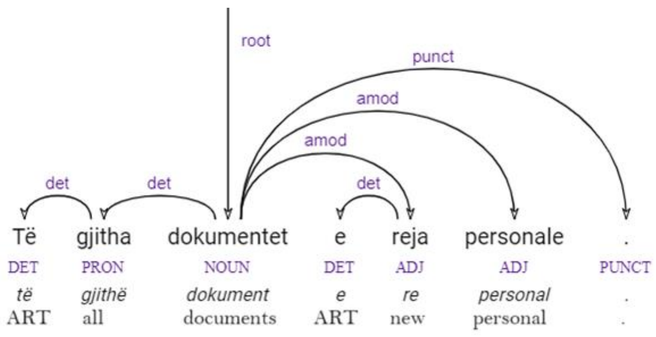
Contributions

Annotation Examples:



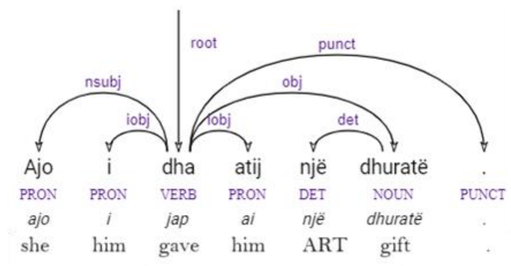
"The flats were covered with thin fir boards or tiles."

Figure 1: Annotated sentence where the root is a verb



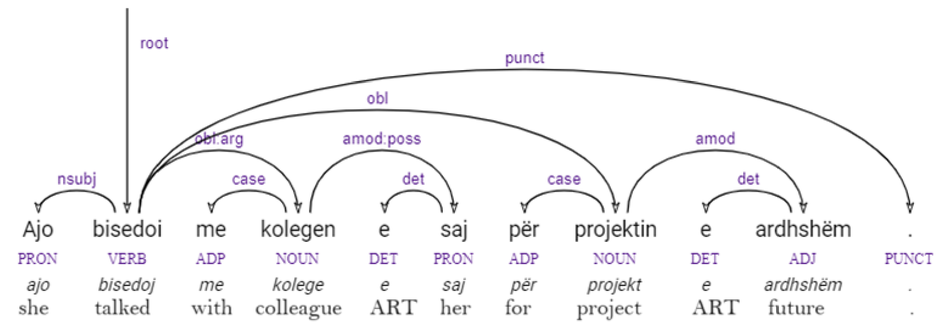
"All new personal documents."

Figure 2: Annotation where the root is a noun



"She gave him a gift."

Figure 3: Example using *iobj* tag



"She talked with her colleague about the next project."

Figure 4: Example using the *obl:arg* tag

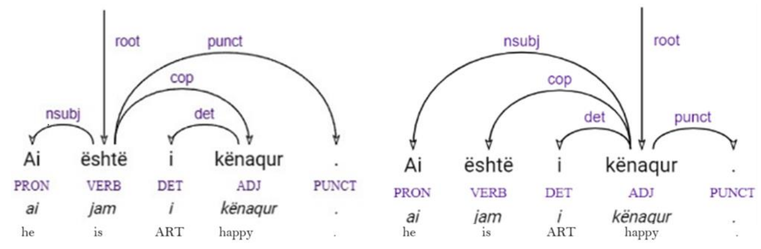
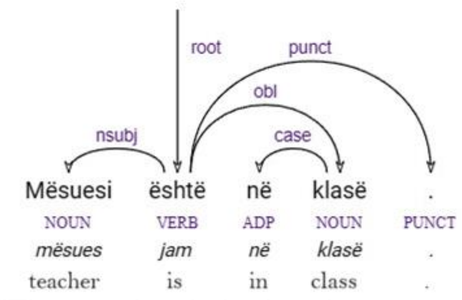
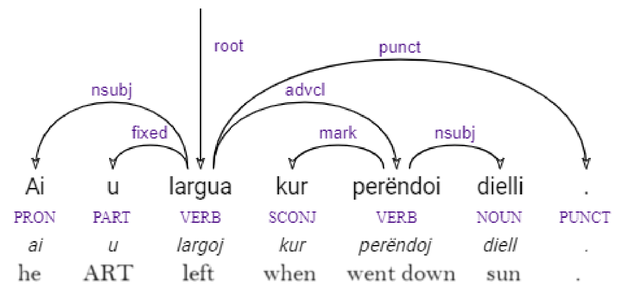


Figure 5: Examples using the *cop* tag



"The teacher is in the classroom."

Figure 6: Example using the *case* tag



"He left at sunset."

Figure 7: Example using *advcl* tag

nkote@fti.edu.al; anila.cepani@unitir.edu.al; alba.haveriku@fti.edu.al

This project was funded by the National Agency for Scientific Research and Innovation as part of the National Research and Development Programs.