Universal Dependencies Treebank for Standard Albanian

Nelda Kote
Polytechnic University of Tirana
Tirana, Albania
nkote@fti.edu.al

Anila Çepani Sema
University of Tirana
Tirana, Albania
anila.cepani@unitir.edu.al

Alba Haveriku
Polytechnic University of Tirana
Tirana, Albania
alba.haveriku@fti.edu.al

Relevant UniDive working groups: WG1, WG3

Abstract

We present the Universal Dependencies treebank for the Standard Albanian language, annotated by linguistic experts with 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. It represents the largest Universal Dependencies treebank available for the Standard Albanian language. In the paper we discuss the selection of sentences for the treebank and addresses key linguistic considerations in adapting the Universal Dependencies framework to align with Standard Albanian grammar. The goal is to contribute to the progress of linguistic analyses and natural language processing in Standard Albanian. The treebank will be made freely available online and in the official Universal Dependencies (UD) repository after the final corrections are completed.

1 Introduction

The Albanian language, which belongs to an isolate branch of the Indo-European family, is the national language of Albania and Kosovo. It is the second national language in North Macedonia and is also spoken worldwide by the Albanian diaspora community. The Albanian language has two main dialects: Gheg, prevalent in northern Albania, and Tosk, used in the southern part of the country and diaspora communities in Greece and Italy. The Standard Albanian language is grounded in the Tosk dialect (Hamp, 2023). In the following sections we discuss the sentence selection methodology and the annotation process. The summary of our contributions is:

- Presenting a new treebank for Standard Albanian language with 25,400 tokens, 21 times larger than the existing treebank TSA (Toska et al., 2020).
- Highlighting the most significant linguistic factors required to align the UD schema with the nuances of the Albanian language, thus creating a valuable resource for interested researchers.

2 The Standard Albanian Treebank

2.1 Data Collection and Selection

The treebank, in total, comprises 4,000 sentences containing 85,000 tokens. To avoid potential conflicts related to proprietary rights, our sentence selection is restricted solely to open corpora. This includes sentences sourced from fiction books, a grammar book, and the Leipzig Corpora Collection (Goldhahn et al., 2012).

Linguistic experts conduct grammatical checks on all selected sentences to identify and rectify errors, a crucial step given that texts in the Albanian language from open-source corpora frequently contain issues like missing letters ("ë" or "ç"), typographical errors, etc.

2.2 The annotation process

The model implemented by Kote et al. (2019) is utilized for pre-annotating for sentence and word segmentation, lemmatization, part-of-speech, and morphological features, followed by a review process to ensure precise sentence and word segmentation, correcting any identified errors. Additional scripts are used to identify and correct remaining errors, including missing morphological features.

Afterward, five linguistic experts manually corrected the incorrect tags and annotated the missing features. Three of the linguistic experts participated only in the morphological annotation, while the other two experts annotated both morphological and syntactic aspects. The syntactic annotation is entirely done manually due to the absence of a trained model for this task.

Throughout the annotation process, two software applications are utilized: the locally installed Conllu Editor (Heinecke, 2019) and the online Arborator Grew (Goldhahn et al., 2012).

The annotation include:

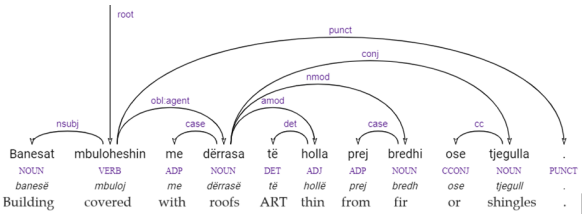
- Sentence segmentation: The selected text is divided into individual sentences, with titles into a text segmented as a separate sentence.
- Words segmentation within a sentence: Sentence segmentation, performed using white space and punctuation marks as boundaries, poses challenges in labeling analytical 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: We apply 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.

2.3 Part-of-Speech and Morphological Annotation

Table 1 shows the list of used tags and their corresponding morphological features.

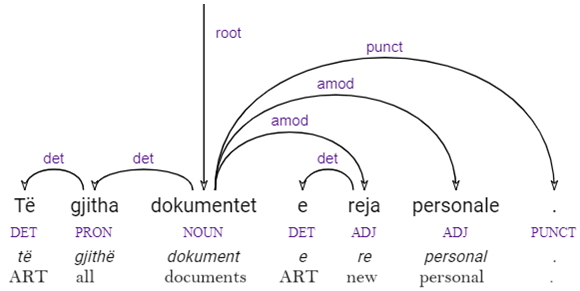
2.4 Syntactic Annotation

Some examples of the syntactic annotation in the Albanian language are represented in the following figures:



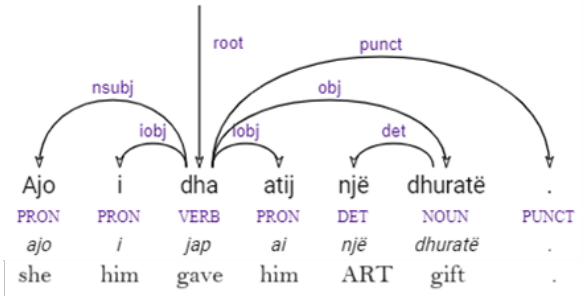
"the flats were covered with thin fir boards or tiles."

Figure 1: Annotated sentence where the root is a verb.



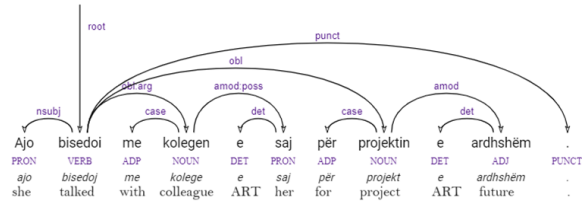
"All new personal documents."

Figure 2: Annotated sentence 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 obl:arg tag.

Despite the Universal Dependencies (UD) emphasis on considering the non-verbal predicate as the syntactic root, our Albanian-specific annotation designates the verb "jam/to be" as the root, which is a copula; differences between UD's copula annotation and ours are illustrated in Figure 5.

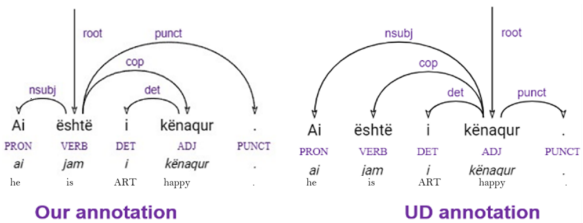
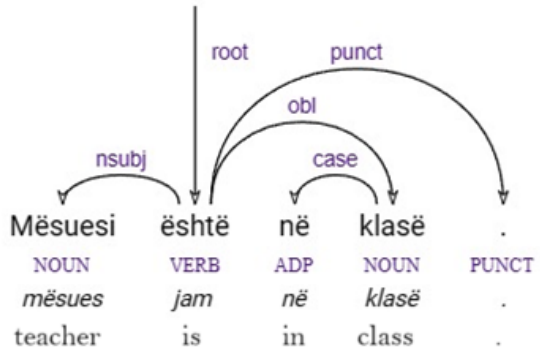


Figure 5: Examples using cop tag.

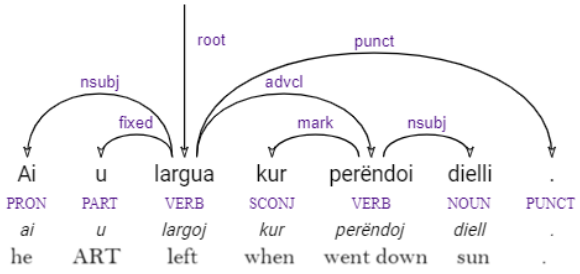
	POS tag	Morphological features
verb	VERB/AUX	mood, time, person, number, voice; (verb form only in case of participle)
noun	NOUN	gender, number, case, definiteness
proper noun	PROPN	gender, number, case, definiteness; (Abbr in case of abbreviation)
adjective	ADJ	gender, number, case, degree
pronoun	PRON	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	-
punctuation marks	PUNCT	-
others	X	Abbr in case of abbreviation

Table 1: The list of the POS tags and morphological features



"The teacher is in the classroom."

Figure 6: Example using case tag.



"He left at sunset."

Figure 7: Example using advcl tag.

3 Conclusions

We present the Universal Dependencies (UD) treebank for the Standard Albanian language, annotated by linguistic experts. To address the annotation challenges specific to Standard Albanian within the UD framework, it is crucial to carefully balance the preservation of the richness of Standard Albanian grammar while mapping the UD tag set and addressing unique language-specific features for a unified annotation. Initially, the UD tag set was aligned with Standard Albanian grammar, and subsequently, linguistic experts manually annotated the corpus.

Acknowledgements

We gratefully acknowledge the support of the National Agency for Scientific Research and Innovation for funding this work under the National Research and Development Programs.

References

ASHSH. 1998, 2002, 2006. *Fjalor i gjuhës së sotme shqipe*. Akademia e Shkencave e Shqipërisë.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*

(LREC'12), pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Eric P. Hamp. 2023. Albanian Languag A.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93, Paris, France. Association for Computational Linguistics.

Nelda Kote, Marenglen Biba, Jenna Kanerva, Samuel Rönnqvist, and Filip Ginter. 2019. Morphological tagging and lemmatization of albanian: A manually annotated corpus and neural models. *CoRR*, abs/1912.00991.

Marsida Toska, Joakim Nivre, and Daniel Zeman. 2020. Universal Dependencies for Albanian. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 178–188, Barcelona, Spain (Online). Association for Computational Linguistics.