Universal Dependencies Treebank for Standard Albanian: A new approach

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

Rozana Rushiti
University of Tirana
Tirana, Albania
rozana.rushiti@unitir.edu.al

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

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

Evis Trandafili
Polytechnic University of Tirana
Tirana, Albania
etrandafili@fti.edu.al

Elinda Kajo Meçe
Polytechnic University of Tirana, Tirana, Albania
ekajo@fti.edu.al

Elsa Skënderi Rakipllari
University of Tirana
Tirana, Albania
elsa.skenderi@unitir.edu.al

Lindita Xhanari
University of Tirana
Tirana, Albania
lindita.latifi@unitir.edu.al

Albana Deda
University of Tirana
Tirana, Albania
albana.deda@unitir.edu.al

Abstract

In this paper, we present a Universal Dependencies (UD) treebank for the Standard Albanian Language (SAL), annotated by expert linguistics supported by information technology professionals. The annotated treebank consists of 24,537 tokens (1,400 sentences) and includes annotation for syntactic dependencies, part-of-speech tags, morphological features, and lemmas. This treebank represents the largest UD treebank available for SAL. In order to overcome annotation challenges in SAL within the UD framework, we delicately balanced the preservation of the richness of SAL grammar while adapting the UD tagset and addressing unique language-specific features for a unified annotation.

We discuss the criteria followed to select the sentences included in the treebank and address the most significant linguistic considerations when adapting the UD framework conform to the grammar of the SAL. Our efforts contribute to the advancement of linguistic analyses and Natural Language Processing (NLP) in the SAL. The treebank will be made available online under an open license so that to provide the possibility for further developments of NLP tools based on the Artificial Intelligence (AI) models for the Albanian language.

Keywords: syntactic dependencies, UPOS, morphological features, Standard Albanian Language, manually annotated corpus

1 Introduction

The Albanian language is part of the Indo-European family and is spoken in Albania, Kosovo, North Macedonia, Montenegro, and other diaspora communities. The language has several unique characteristics that distinguish it from other European languages. There are regional variations and dialects in the language, with the two main dialects being Gheg, which is spoken in northern Albania, and Tosk, which is spoken in southern Albania, as well as in diaspora communities in Greece and Italy. These two dialects have distinct lexical differences. The Standard Albanian Language (SAL) is based on the Tosk dialect (Hamp, 2023).

Universal Dependencies (UD) (Nivre et al., 2020) is a framework focused on the provision of research at a multilingual level for morphological and syntactic annotation. The currently available treebank for Albanian language in the UD framework consists of only 60 sentences, annotated with lemmas, morphological, and syntactic features (Toska et al., 2020).

This paper presents the Standard Albanian Language Treebank (SALT) annotated conform UD framework by expert linguistics with the support of information technology professionals, consisting of 24,537 tokens (1,400 sentences). The annotation includes sentence segmentation, word segmentation, universal part-of-speech (UPOS) tags, morphological and syntactic features, and lemmas, offering a new valuable resource for the study of the SAL.

Contributions: We discuss the methodology to select the sentences of our treebank and the inclusive criteria. Furthermore, after extensive work to identify special linguistic features in SAL, we present the decisions made by the expert linguistics

group for aspects such as verb form, noun declension, adjective agreement, and different syntactic problems, essential to create a proper standard for the annotation. The summary of our contributions is:

- Presenting the new SALT treebank with 24,537 tokens, 21 times larger than the existing treebank TSA (Toska et al., 2020).
- Emphasizing the most significant linguistic characteristics required to align the UD schema with the characteristics of the SAL by creating a valuable resource for interested researchers.

The rest of this paper is organized as follows: In Section 2, we review significant background and related research works. In Section 3, we discuss the treebank development, sentence collection, and the annotation process. Section 4 covers the treebank annotation schema and related discussions. Finally, in Section 5, we conclude our work and discuss directions for future research.

2 Background and Related Work

Research efforts for low-resource languages like Albanian have been historically constrained. Although there have been several attempts to develop annotated corpora in the Albanian language, they have either remained closed-source or proved to be too limited in size and lacking interconnection between them.

The largest existing corpora for the Albanian language are the "Albanian language corpus" (16.6 million tokens) created by the Saint-Petersburg Institution (Arkhangelskij et al., 2011) and the sqGlobe corpus (1 million words) by the Beijing Foreign Studies University (Ke et al., 2012) annotated with POS tags and lemmas. These two corpora are the largest ones for the Albanian language, but due to them being closed-source, they are not suitable for further research works. Meanwhile, Caka and Caka (2011) have created a closed-source corpus with one million words, which is lemmatized and includes grammatical properties.

Kote et al. (2019) present a corpus containing 118,000 tokens, annotated at the morphological level based on the UD schema, the UniMorph project (Kirov et al., 2018) present a treebank with 33,483-word forms and 589 lemmas and Toska et al. (2020) present the first official UD treebank for the Albanian language but containing only 60 sentences. Other related works include Kadriu (2013), Kabashi and Proisl (2018), Misini et al. (2020), Ebert et al. (2022) and Mati et al. (2021) that have contributed in different tagsets and small treebank for standard and Gheg Albanian.

3 Standard Albanian Language Treebank

This section outlines the development of the SALT treebank. Two expert linguistic conducted manual annotation due to the lack of preprocessing tools and resources for SAL. Tables 1 and 2 show statistics about the SALT treebank.

Number of sentences	1,400
Number of tokens	24,537
Multiword tokens	87
Avg. sentence length	18

Table 1: Statistics of the SALT treebank.

UPOS	**frequency**	**Deprel**	**frequency**
NOUN	5,353	punct	2,893
PUNCT	2,908	det	2,304
DET	2,736	case	2,099
VERB	2,697	advmod	1,647
ADP	2,139	nsubj	1,620
PRON	1,738	nmod	1,467

Table 2: Frequency of the most used tags.

3.1 Data Collection and Selection

The treebank consists of 1,400 sentences containing 24,537 tokens. To prevent potential proprietary rights conflicts, we selected sentences from open corpora. The sentences are extracted from fiction books, a grammar book, and the Leipzig Corpora Collection (Goldhahn et al., 2012). Before annotation, all the sentences are grammatically corrected for any error by the expert linguistics. This step is necessary because texts in the Albanian language available in open-source corpora often exhibit grammatical errors such as missing letters like "ë" or "ç", typographical mistakes, etc.

3.2 The Annotation Process

To facilitate the annotation process the selected sentences are pre-annotated using the model proposed by Kote et al. (2019) for segmentation, lemmatization, part-of-speech, and morphological features. Subsequently, the annotated sentences are reviewed

	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 3: The list of the POS tags and morphological features.

to ensure accurate sentence and word segmentation, with any errors corrected as needed. Additional scripts are used to identify and correct other errors and add missing morphological features. Following this, expert linguistics manually reviewed and annotated all the sentences. The syntactic annotation is entirely done manually, as there is no available trained model for it. The expert linguistics have reached an agreement among themselves to use a standardized method for annotation.

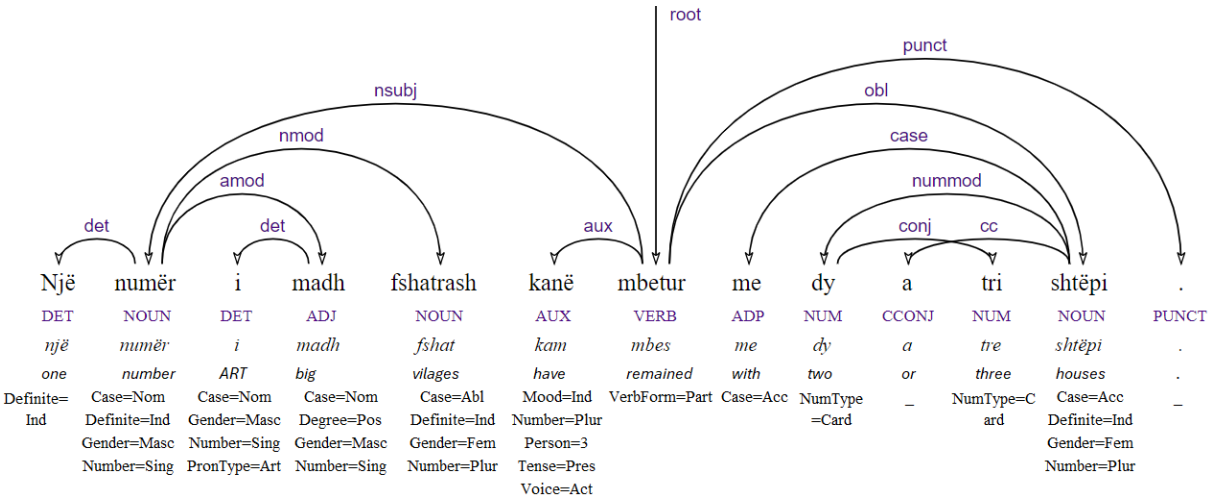
Two software applications, Conllu Editor (Heinecke, 2019) and Arborator Grew (Guibon et al., 2020), are used for annotation.

The annotation include:

- **Sentence segmentation:** The selected text is segmented in sentences, with titles segmented as a separate sentence.
- **Word segmentation:** Word segmentation was performed using white space and punctuation marks as boundaries, leading to challenges in identifying analytical grammatical forms and various expressions. As the Albanian language is a synthetic-analytical language, with a prevalence of synthetic features but a tendency towards analytic structures (compound verb tenses, the future tense, certain verb forms, nonfinite verbs, the nominative adjective, genitive of nouns, conjunction locutions, prepositions, and phraseological expressions), labeling becomes intricate.
- **Lemmatization:** Expert linguistics used the Albanian National Dictionary, (ASHSH, 1998, 2002, 2006) to determine the lemma of a word. The lemma is assignment based on the context and meaning of the word form in the sentence.
- **Part-of-speech tags:** A total of 17 part-of-speech tags from the UD tagset are utilized.
- **Morphological features:** We applied corresponding morphology features based on the word's part-of-speech tag.
- **Syntactic annotation:** A total of 32 syntactic tags from the UD tagset are utilized.

4 Annotation Schema

In this section, we discuss the key considerations of the annotation schema used to annotate text data within the UD framework. The grammar of the Albanian language has a complex inflection schema and a rich morphological and syntactic structure, which presents several challenges in annotation due to the presence of unique features specific to the language.



"A large number of villages are left with two or three houses."

Figure 1: Annotation of a sentence with compound verb tense.

In the annotation process, the grammatical guidelines published by the Albanian Academy of Sciences (Agalliu et al., 2002) are taken into account.

4.1 Part-of-Speech and Morphological Annotation

We have utilized 10 universal part-of-speech tags (verb, noun, adjective, pronoun, adverb, numeral, interjection, preposition, particle, conjunction) from the UD framework, along with their respective grammatical features. Furthermore, we have utilized 4 other tags for various elements present in SAL grammar, such as articles, abbreviations, symbols, and punctuation marks. Table 3 shows the list of used UPOS tags and their corresponding morphological features.

The verb system is one of the most complex aspects of SAL grammar, comprising 6 moods and 14 tenses in total (Agalliu et al., 2002). However, when annotating each word separately, some moods and tenses for compound verb tenses may not be explicitly displayed, such as the future indicative tense, future perfect tense, etc. Each part of the compound verb is separately annotated depending on its form. Another case to discuss is the annotation of the verbs "kam/to have" and "jam/to be," that are used for three different purposes:

- As copula, it is annotated with VERB tag.
- As auxiliary verb to form the compound tenses, it is annotated with AUX tag.
- As main verb, it is annotated with VERB tag.

Unfortunately, there is no specific morphological tag available to distinguish between the uses of this verb. This necessitates defining these cases as verbs with their full lexical meaning, even when they function as auxiliary verbs. Figure 1 illustrates an example.

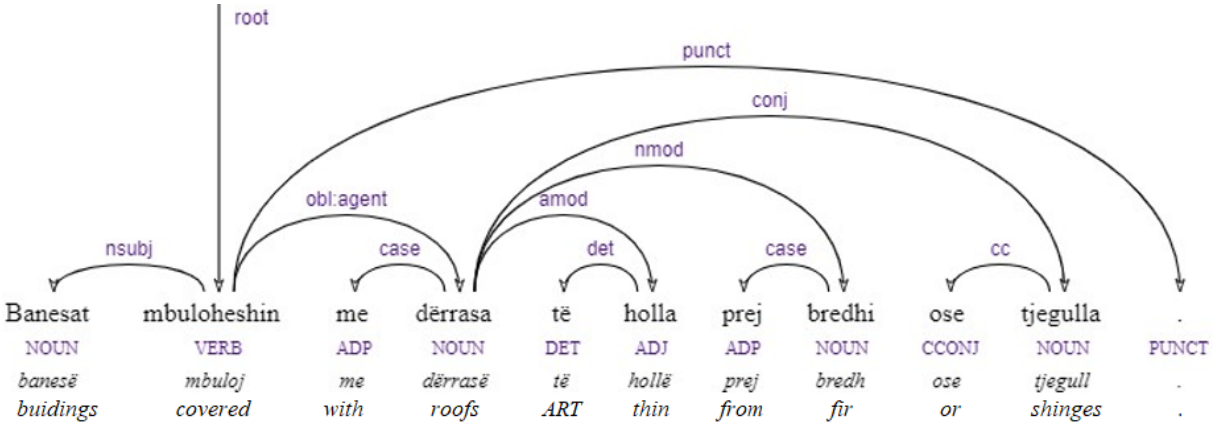
The adjectives agree with the governing nouns in number, gender, and case regardless of the degree. There are three degree categories: positive, comparative, and superlative. The comparative and superlative forms are created as analytic forms utilizing lexical elements, with the positive degree being the main focus of annotation, as it serves as the base for these forms. Another aspect not covered in adjective annotation is the categorization of articulated and non-articulated adjectives. The article of the articulated adjective is annotated as DET, like the adjective "i madh/big" in the sentence presented in Figure 1.

Pronouns are classified into seven distinct classes, each annotated with different morphological features. Some share common attributes such as case, number, and gender, while others lack specific categories, including abbreviations that are also annotated as pronouns.

The conjunctions and adverbs are annotated based on their types, but the annotation for conjunctions doesn't encompass semantically related subtypes.

Prepositions are annotated with case morphological features to aid in syntactic analysis and to determine the type of syntactic relationships formed with prepositional phrases.

Given the diverse and multifaceted nature of the



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

Figure 2: Annotated sentence where the root is a verb.

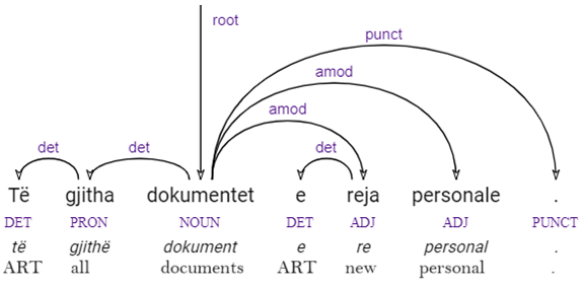
particles, they are not annotated with specific type tags, as they encompass both semantic and grammatical dimensions.

It is important to clarify how we have used the determiner (DET) in our annotation based on the UD framework. In traditional Albanian grammar, articles are not categorized as a separate part of speech "Determiner - DET" but are instead treated as morphological elements associated with nouns, adjectives, and pronouns, inheriting features such as gender, number, and case. Despite this, we have aligned articles with the DET category in UD based on their functional roles, which include indicating definiteness, gender, number, and case. This mapping preserves the grammatical features of articles while ensuring consistency with UD principles. By doing so, we maintain the traditional grammatical structure of Albanian while leveraging UD's universal annotation schema, thereby providing an accurate and comprehensive representation of Albanian articles within the UD framework.

4.2 Syntactic Annotation

The root of the sentence is indicated by the root tag, usually by labeling the principal verb, sentence designer, or principal unit verb in compound sentences. When the verb (predicate), which marks thematic roles in a sentence, is absent (due to ellipses), and multiple orphaned subordinates exist, by agreement we decide that one of these subordinates takes on the role of the root while the others relate to it. As a result, a noun is labeled as the root, although in Albanian language, an adjective or another noun can also assume the root role through agreement. Figure 1, Figure 2, and Figure 3 show examples of annotated sentences with a verb as the root and a noun as the root.

The nsubj (nominal subject) tag is the external argument (the headword) or syntactic subject representing the agent acting, whether expressed through a noun as shown in Figure 1 and Figure 2, pronoun, numeral, or nominal expression.

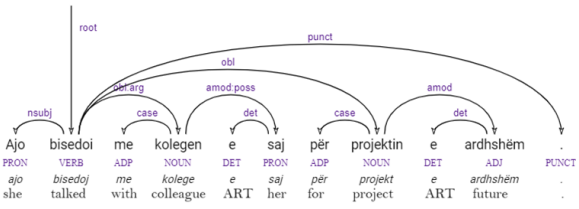


"All new personal documents."

Figure 3: Annotated sentence where the root is a noun.

The obl (nominal oblique) tag is used for a noun phrase, specifically when its head is a nominative case preposition+noun, pronoun, or a noun phrase, or any preposition in another case, serving as a non-core or complement argument. Often, this functions like an adverb linked to another verb, adjective, or adverb. To specify the noun subject of a passive verb, the subtype pass (passive) is typically used in the UD framework, but during our labeling, we employed obl: agent for such structures. The obl:agent is used for an oblique noun phrase indicating the agent, and the obl: arg is used for an oblique noun phrase functioning as an argument, as in the example shown in Figure 4.

The obj (direct object) tag is used for direct object of the verb, where the verb action falls, irrespective of whether it's a noun or a pronoun (full



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

Figure 4: Example using obl:arg tag.

form+clitic/full form/clitic) in the accusative case. It happens in languages where obj is labeled with the morphological case. Figure 5 shows examples.

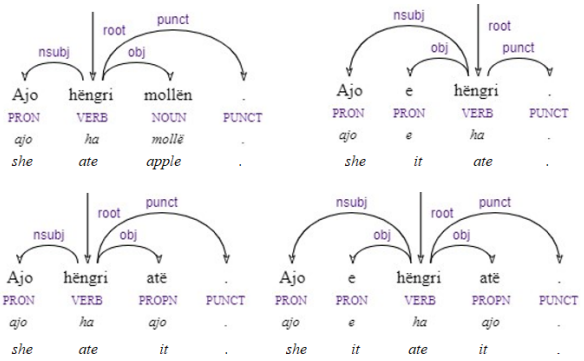
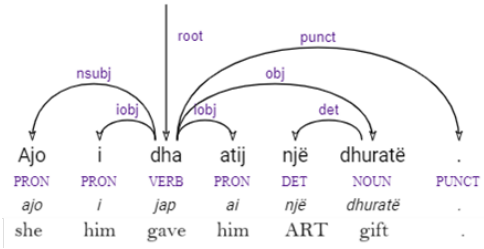


Figure 5: Examples using obj tag.

In the Albanian language, iobj (indirect object) tag is associated with arguments in the dative case, such as in the example shown in Figure 6.



"She gave him a gift."

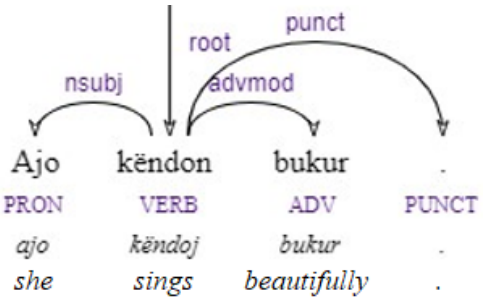
Figure 6: Example using iobj tag.

The nmod (nominal modifier) tag represents a nominal modifier that modifies a noun and we use it for noun dependents of another phrase, such as attributes or complements, associated with the head of the noun phrase.

The amod (adjectival modifier) tag is used not only for adjectival modifiers of a noun or a pronoun but also pronominal modifiers to which the poss subcategory has been added when they are possessive pronouns.

The appos (appositional modifier) tag is used for a noun or noun phrase that explains or defines another noun in the role of an affix. It is also used to link names or noun phrases when providing supplementary information such as email addresses, phone numbers, or residential addresses.

The advmod (adverbial modifier) tag is used for adverb or adverbial phrase emphasizes the modification of another verb, adjective, or adverb, as in the example shown in Figure 7.



"She sings beautifully."

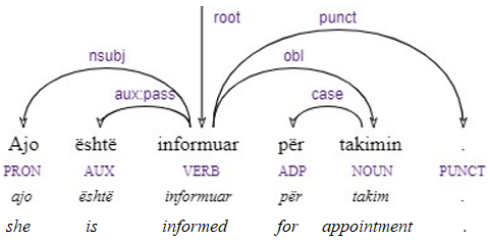
Figure 7: Example using advmod tag.

In Albanian language, a limited set of adverbs can also modify nouns, as in "vetëm të hënën/only on Monday", where the advmod tag is used, emphasizing the adverb's role, like advmod:emph.

The term "adverbial modifier" encompasses compounds functioning as adverbs, whether adverbs, non-clausal phrases, or nouns in specific morphological cases, as is the case in the Albanian language. We distinguish modifiers as adverbs (advmod) and others as non-clausal phrases or adverbs (obl). However, we do not differentiate between predicate verb modifiers, so adverbs in a strict sense, and modifiers of other modifying words like adjectives or adverbs, as all these are under the advmod category. The obl tag is used when the circumstantial element is obligatory (an argument), while the ad mod tag is used when the circumstantial element is optional.

The aux (auxiliary) tag indicates a verb that is linked to another verb (predicate), typically serving as an auxiliary verb used to form analytical verb forms as shown in Figure 1. It also includes semi-auxiliary verbs with modal significance, which express the manner of an action and can often have full lexical meanings, "Ajo nisi/mund të kryejë detyrat e shtëpisë / She begin to/can do her homeworks". The aux tag is utilized to represent passive voice constructions and, in languages with a grammaticalized (periphrastic) passive form, the subtype

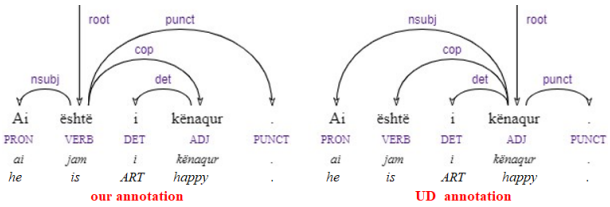
aux:pass is encouraged for usage. So, the aux: pass (passive auxiliary) tag is used for auxiliary verbs associated with the past participle, as observed in the sentence shown in Figure 8.



"She has been informed about the appointment."

Figure 8: Example using aux: pass tag.

The cop (copula) tag represents the relationship between a subject and a non-verbal predicate, connecting the noun with the subject. In UD, the non-verbal predicate is the root from which all other syntactic connections are created, but referring to the specifics of Albanian, we have annotated as root the verb "jam/to be" which is a copula. Figure 9 shows the difference between the UD annotation of the copula and our annotation.



"He is happy."

Figure 9: Examples using cop tag.

The nummod (numeric modifier) tag is used for a numerical modifier of a noun, which can be any number that modifies and indicates its quantity. A number that serves as a label for another entity and does not indicate quantity is not labeled as nummod but nmod as in: "Takimi do të jetë në dhomën 4./The meeting will be in room 4.".

The det (determiner) tag is used for a determiner that specifies the noun it modifies. It connects a noun/adjective head, as well as certain types of pronouns with their determiner or modifier. In the context of SAL, this tag applies to all non-significant determiners of the name, all word-forming and shape-forming determine, and so in addition to the det tag, there must be subcategories like det: adj, det: pron, and det: poss.

The case (case marking) tag is used for the analytical case marker, which is treated as a special syntactic word, such as prepositions in SAL as shown in Figure 1.

The conj (conjunct) tag indicates connections with coordination between members and parts of the compound sentence. The main part of the connection is called the first part, and all subsequent parts are connected to it via the conj tag.

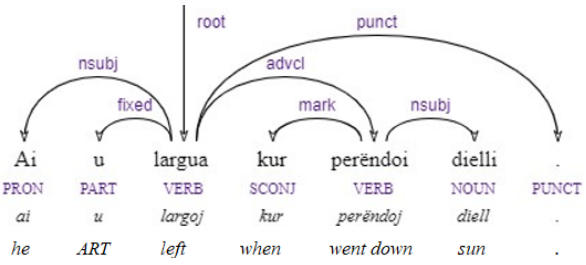
The cc (coordinating conjunction) tag connects a coordinating conjunction to a word-member or compound sentence with coordination.

The csubj (clausal subject) tag is used in the subject sentence.

The xcomp (open clausal complement) tag is used for predicate clauses, and in the case of complementary clauses, the ccomp (clausal complement) tag is used.

The acl (clausal modifier of a noun) and acl:relcl (relative clause modifier) are used for a dependent and relative clause that modifies nouns.

The advcl (adverbial clause modifier) tag is used to indicate a dependent clause that functions as an adverbial modifier. Figure 10 shows an example.



"He left when the sun went down."

Figure 10: Example using advcl tag.

We should emphasize that tags such as csubj, ccomp, xcomp, acl, and advcl, which are used to connect in minimal structures, the leading verb with a dependent unit or a noun with a dependent unit, referring to the classificatory criterion of which component of the leading unit the subordinate unit modifies, so they are labeled as either complementary, necessary dependents (csubj, ccomp, xcomp) or as optional determiner-relational dependents modifying a noun in the governing unit acl, or as additional adverbial circumstantial dependents advcl.

The discourse tag is used when an element of discourse is in a sentence, such as an exclamation that shows emotional content.

The dislocated tag is used for an element that appears separately from its syntactic position, and also intermediate words or phrases, interlaced, and

sentences that have no grammatical connection with the rest of the sentence.

Adverbial, prepositional, and conjunctional locutions are annotated using the fixed tag that expresses fixed expression.

A fixed phrase composed of two or more words that convey a particular grammatical structure is annotated using the fixed:form tag.

The flat (flat multiword expression) tag is used for phraseological expressions or various nouns.

The list tag is used when parallel elements are listed syntactically. In the case of homogeneous members with coordinating conjunctions, the conj tag is used. Meanwhile, the list tag can be used only when there are no conjunctions.

The parataxis tag refers to phrases or sentences which lack conjunctions or connecting elements.

The mark (marker) tag is used for subordinating conjunctions, particles, etc.

The punct (punctuation) tag is used for punctuation marks. Since they do not follow a typical dependency relationship, several criteria are used to determine their main associated word. Generally, periods, exclamation points, and question marks, which indicate sentence conclusions, are connected to the main verb serving as the root. Commas are connected to coordinating or subordinating conjunctions, or words designated as part of a list. Paired punctuation marks, such as quotation marks, brackets, and sometimes hyphens, are connected to the same word.

The vocative tag indicates an element in the sentence that is a noun used to address a person.

The orphan tag expresses dependencies in the case of a missing head.

4.3 Annotation Discussions

Acknowledging the complexity of defining word boundaries in SAL due to its various grammatical structures, including compound verb tenses and phraseological expressions, we use the ”fixed:” relation of UD for lexicalized multi-word expressions like the subjunctive mood form ”të lexoj”. However, we encountered challenges with the particle ”të,” especially when it merges with the accusative short form ”e,” forming ”ta (të + e) lexoj”. This merging complicates the use of the ”fixed:” relation, as the resulting form must be accurately represented in the annotation. Therefore, while we have applied the ”fixed:” relation where appropriate, addressing the merging of particles requires additional scripts or manual intervention to ensure precise tagging of these compound structures.

Determining the lemma of adjectives can be challenging because, depending on the noun they modify, they can exist in both feminine and masculine forms. In these cases, we have chosen to use the masculine form specified in the dictionary as the lemma (e.g., “mahnitshëm/amazing” is the lemma used even for the masculine form “i mahnitshëm” and feminine form “e mahnitshme”).

Articles, pronouns, and abbreviations can appear in different forms depending on the words they modify. In this case, each distinct variant found in the dictionary should be a lemma to ensure a precise and unambiguous representation with accurate labeling for these linguistic components.

Different elements, including subordinating conjunctions and various participles, are labeled using a mark tag. However, there is no specific label for passive particle and particle ”të” that form the relative mood, the future tense, and the affirmative and negative participles.

The nsubj tag is used for all types of headwords (expressed by noun, pronoun, phrase) without the option to use additional tags for subcategorization.

There is no additional syntactic relation label for verbs that lack full lexical meaning and necessitate a complementary predicate. In this case, cop tag is used similar to the verb “jam/to be” in the nominal predicate. Indeed, the presence or absence of this sentence’s element is a discussion in SAL syntax, much like the acceptance of the nominal predicate itself. Even for modal verbs (e.g., “mund/can”, “duhet/should”) and aspectual verbs expressing the initiation, continuation, or completion of actions (e.g., “filloj/start”, “vazhdoj/continue”), there are no dedicated labels. In these cases, we use the aux relation used for the auxiliary verbs “jam/to be” and “kam/to have”.

The oblique nominal with prepositions remains a subject of controversy among researchers and linguists, even to this day. As UD does not have a distinct label, we used obl tag.

No specific label to distinguish pronouns before or after a noun. For example, to label indefinite pronouns like “asnjë/none”, “pak/few”, “shumë/many” before a noun, the nummod tag is used because they express indefinite quantities and are treated as quantifiers, but this tag is not used when they follow a noun.

Conjunctions formed by a noun with a posses-

sive pronoun are labeled with amod, similar to the conjunctions formed by a noun with an adjective by adding a subcategory poss, as amod:poss. This tag is used even for indeclinable pronouns, thus broadening the usage of this label. Also, an issue to discuss is the textual conjunctions that appear at the beginning of a sentence, intermediate words, and interjections. The question is: "How should they be integrated into the sentence structure?".

For various types of punctuation marks (period, comma, question mark, exclamation mark, quotation marks, brackets, etc.), there is no specific label or subcategorization to define their functions. Therefore, they are all labeled with punct tag.

5 Conclusion and Future Work

This paper presents the Standard Albanian Language Treebank (SALT), the first Universal Dependencies (UD) treebank for the Standard Albanian Language (SAL), annotated by expert linguists. SALT includes annotations for syntactic dependencies, part-of-speech tags, morphological features, and lemmas. Adapting the rich and complex grammar of the Albanian language to the UD schema involves significant challenges, such as the absence of direct mappings and the ambiguity in assigning appropriate tags.

We present an overview of the language's grammatical structure, providing a detailed analysis of its key linguistic features. Additionally, we discuss methods for annotating texts in SAL according to the UD framework. Overcoming annotation challenges requires delicately balancing and harmonizing the richness of the language's grammar with adaptation to the UD tagset, while addressing the unique language-specific features for unified annotation. Expert linguists initially mapped the UD tagset to align with the language's grammar and subsequently performed manual annotations on the treebank.

As future work, we aim to use our proposed treebank, composed of 1,400 sentences, as a training and testing dataset for an Albanian language parser. This automated tool will facilitate the annotation of a larger treebank, aiding linguistics, computer scientists and related fields to conduct further research work on the Albanian language.

Our treebank will be available online with open access for research purposes, aiming to foster advancements in NLP research for SAL.

Acknowledgements

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

References

F. Agalliu, E. Angoni, S. Demiraj, and e. al. 2002. *Gramatika e gjuhës shqipe*. Instituti i Gjuhësisë dhe i Letërsisë (Akademia e Shkencave e RSH), Tirana.

T. Arkhangelskij, M. Daniel, M. Morozova, and Rusakov A. 2011. Korpusi i gjuhës shqipe: drejtimet kryesore të punës. In *Proceedings of Shqipja dhe gjuhët e Ballkanit — Albanian and Balkan Languages.*, page 635–642, Prishtina, Kosovo. ASHAK.

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

Nebi Caka and Ali Caka. 2011. Korpusi i gjuhës shqipe – rezultatet e para, problemet dhe detyrat.

Christian Ebert, Adrian Kuqi, Paul Widmer, and Barbara Sonnenhauser. 2022. Ud gheg pear stories: An annotated treebank of gheg albanian as spoken in switzerland. *PREPRINT*, Version 1.

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).

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Eric P. Hamp. 2023. Albanian Language.

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.

Besim Kabashi and Thomas Proisl. 2018. Albanian part-of-speech tagging: Gold standard and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Arbana Kadriu. 2013. Nltk tagger for albanian using iterative approach. *Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces*, pages 283–288.

J. Ke, Q. Jin, S. You, T. Han, Y. Feng, X. Wang, Z. Hu, E. Laçi, E. Allmetaj, T. Chen, W. Zhang, H. Zhang, Y. Lu, and W. Ai. 2012. The sqglobe corpus (a balanced corpus of 1m-word contemporary written albanian, lemmatised and pos-tagged).

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Diellza Nagavci Mati, Mentor Hamiti, and Elissa Mollakuqe. 2021. Morphological tagging and lemmatization in the albanian language. *SEEU Review*, 16(2):3–16.

Arta Misini, Ercan Canhasi, and Samedin Krrabaj. 2020. Albanian syntactic parsing. *ICT Innovations 2020, Web Proceedings ISSN null*, pages 135–150.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.