Universal Dependencies for Albanian

Marsida Toska and **Joakim Nivre**
Uppsala University
Department of Linguistics and Philology
Uppsala, Sweden
marsida.toska.1494@student.uu.se
joakim.nivre@lingfil.uu.se

Daniel Zeman
Charles University
Faculty of Mathematics and Physics
Prague, Czechia
zeman@ufal.mff.cuni.cz

Abstract

In this paper, we introduce the first Universal Dependencies (UD) treebank for standard Albanian, consisting of 60 sentences collected from the Albanian Wikipedia, annotated with lemmas, universal part-of-speech tags, morphological features and syntactic dependencies. In addition to presenting the treebank itself, we discuss a selection of linguistic constructions in Albanian whose analysis in UD is not self-evident, including core arguments and the status of indirect objects, pronominal clitics, genitive constructions, prearticulated adjectives, and modal verbs.

1 Introduction

Albanian is an Indo-European language and also part of the Balkan Sprachbund.[1] It is spoken primarily in Albania and secondarily in neighbouring Balkan countries by Albanian minorities but also elsewhere in Europe and outside by the Albanian diaspora. Its vocabulary, but even more so its grammar, exhibits a lot of similarities and parallelisms with the languages of the Balkan Sprachbund, while it also features linguistic constructions that could be characterized as idiosyncratic.

Computational and other online resources that could facilitate NLP research and comparative studies of Albanian are scarce. A large part-of-speech tagged corpus for the language was created only recently by Kote et al. (2019), but there is still no corresponding treebank with full syntactic annotation, which means that it is hard to develop language technology applications that require both tagging and parsing. It is in this context that we have developed the first Universal Dependencies (UD) treebank for Albanian, called UD Albanian-TSA.[2] Although still very limited in size, it constitutes a first step towards developing a large-scale treebank within the UD scheme, enabling NLP research as well as comparative studies involving other languages, and there is also research showing that even a few annotated sentences can contribute to good parsing results (Meechan-Maddon and Nivre, 2019).

In the following sections we introduce some of the key features of the Albanian language (Section 2), provide a brief summary of related work with regard to NLP for Albanian (Section 3), and describe the steps taken to develop the treebank (Section 4). We then discuss in some detail a selection of linguistic constructions in Albanian that pose challenges for the UD annotation framework and that are interesting from a cross-linguistic perspective (Section 5).

2 The Albanian Language

Albanian belongs to the Indo-European family of languages, but it constitutes its own branch within the family. It is spoken by around 7.5 million people, of which, due to the Albanian diaspora, less than 3 million are estimated to reside in Albania (Hoxha and Baxhaku, 2019). There are two main Albanian dialects, the Tosk and the Gheg, with the former being the one Standard Albanian is based on. Its alphabet relies on the Latin one and comprises 36 letters, 9 of which are digraphs (dh, gj, ll, nj,

[1] Lindstedt (2000) explains that the Balkan Sprachbund comprises Albanian, Greek, Balkan Slavic, Balkan Romance and Balkan Romani languages.

[2] TSA is short for Treebank for Standard Albanian.

rr, sh, th, xh, zh) while 2 have diacritics (ë, ç) (Karanikolas, 2009). The dominant word order is SVO, but the rich morphology often allows relatively free word order. Breu (2010) lists the following salient characteristics of the Albanian language, which also apply, with some variation, to most languages of the Balkan Sprachbund:

- Lack of infinitive
- Analytical comparison system:

  (1) më i miri
  PART ART.M good.M
  'the best'

- Future construction with "will/want":

  (2) do të vij
  will to come.1SG.PRES
  'I will come'

- Object redoubling in dative and accusative (clitic doubling):

  (3) i-a dhash libr-in shok-ut
  him.CL.DAT-it.CL.ACC gave.1SG.PAST book-the.ACC friend-the.DAT
  'I gave the book to the friend'

- Suffixed definite article:

  (4) libr-in
  book-the
  'the book'

- Complex verb system (3 tenses, 6 moods, etc.), e.g., admirative mood:[3]

  (5) qenke i shpejtë
  be.2SG.PRES.ADM ART fast
  'You are surprisingly/unexpectedly fast'

3 NLP for Albanian

In recent years, the development of NLP resources for Albanian has been increasing. Most research has focused on the morphological analysis and the creation of part-of-speech tagging models. Trommer and Kallulli (2004) first introduced a morphological analyzer that made use of off-line components, while later on Piton and Lagji (2008), performing a morphological study of Albanian, developed electronic dictionaries of inflected forms as well as transducers using NooJ. In the area of part-of-speech tagging, Kabashi and Proisl (2018) proposed a part-of-speech tagset after noticing there did not exist one of moderate size, mapping it also to the UD tagset (Nivre et al., 2016; Nivre et al., 2020). They built on their own previous work (Kabashi and Proisl, 2016) and that of other researchers, such as that of Hasanaj (2009), who had developed a statistical part-of-speech tagging model with accuracy around 70%, and Kadriu (2013) who had presented a part-of-speech tagging model using the NLTK toolkit.

In 2019, the authors[4] of the Albanian National Corpus[5] (Morozova and Rusakov, 2014) (ANC) developed and made publicly available in the official website of the ANC a lemmatizer, tagger and morphological analyzer for Albanian. However, they use their own part-of-speech tagset and do not rely on the UD annotation scheme, and there is no disambiguation of either lemmatization, tagging or morphological analysis. ANC itself contains around 20 million tokens and is therefore the largest resource currently available for Albanian.

Also recently, Kote et al. (2019) presented an Albanian corpus with part-of-speech tags and morphological features containing around 118,000 tokens. Additionally, the team trained a neural morphological

[3]The admirative mood in Albanian incorporates a number of modality nuances such as admiration, surprise etc.
[4]Maria Morozova, Alexander Rusakov, Timofey Arkhangelsky
[5]http://albanian.web-corpora.net/

tagger and lemmatizer which achieved promising results, the best of which being 92.74% in part-of-speech tagging. The annotation of the corpus was based on the UD guidelines and underwent a manual review.

4 Treebank Development

In this section, we describe the development of UD Albanian-TSA. Given the lack of preprocessing tools and resources for Albanian compatible with the UD framework,[6] most tasks undertaken in the creation of the treebank were performed manually. However, some steps, such as word segmentation, lemmatization, tagging and morphological analysis were semi-automated through scripts that we developed or tools that were available from other researchers, as described in more detail below. In the end, the entire treebank underwent manual checking and correction, so as to resolve cases of ambiguity, eliminate errors and ensure overall consistency, and was also tested with the UD validation script.

4.1 Data Selection

Although our initial intention was to work with data from the ANC, which contains a wide compilation of texts from different genres, we eventually collected our data from random entries of the Albanian Wikipedia because of the free license. The selection of the sentences was manual and guided by the aim to include as many linguistically diverse structures as possible. Our data set consists of 60 sentences in total corresponding to 922 tokens.

4.2 Word Segmentation and Lemmatization

The segmentation of sentences into words was performed based on white-space delimiters and punctuation. This means, for instance, that the adjective *i zi* (black), which is a so-called prearticulated adjective, was split into the preposed article *i* and the main word *zi*, despite the latter being ungrammatical and not falling within a word class on its own.[7]

For lemmatization, we used the lemmatizer developed by the Albanian National Corpus team.[8] However, since this lemmatizer does not disambiguate between identical tokens with different lemmas, manual disambiguation was required. For example, the token *vinte* (3rd person singular past tense verb), depending on the context, could be lemmatized as either *vij* (to come) or *vë* (to put).

4.3 Morphological Features

The morphological analysis of the tokens was to a great extent manual, except for cases where it was possible to automate the process through scripts. The classes that were assigned morphological features are the following: verbs, nouns, adjectives, pronouns and determiners.

For verbs, we included features such as aspect, mood, number, person, tense and voice, unless these occurred in form of participles or gerunds, in which case only the feature `VerbForm` was used. For nouns, we decided to include the following features: case, definiteness, gender and number. In addition, the feature `NounType=Het` was added for nouns displaying different gender in singular and plural (also known as "two-gender nouns" or "dual nouns").[9] The features case, gender and number were also adopted for pronouns; in addition, the type of pronoun is specified by the feature `PronType`. For adjectives, we narrowed down the number of features to two, namely gender and number, unless further features were explicitly marked, such as `AdjType` (if the adjective was grammatically a perfect passive participle). The last class to get features was that of determiners, for which only the gender was specified.

[6] The corpus and tools developed by Kote et al. (2019) unfortunately appeared only after our work had been finished.

[7] The syntactic analysis of prearticulated adjectives is discussed in detail in Section 5.4.

[8] `https://bitbucket.org/timarkh/uniparser-albanian-grammar/src/master/`

[9] Beqiraj (2014) states that in Albanian, two-gender nouns are masculine in the singular (originally classified as neuter) and change to feminine in the plural, usually by adding the suffix -e or -ra. He also mentions that this phenomenon appears (although less frequently and systematically) in other Indo-European languages as well, such as French, e.g., l'amour (singular masculine, "the love") vs. les amours (plural feminine, "the loves").

4.4 Part-of-Speech Tagging

In order to speed up the process of assigning UD part-of-speech tags to words, we created scripts that tagged closed-class words such as particles, adpositions, determiners, coordinating and subordinating conjunctions, numerals, punctuation marks, interjections and symbols. As for verbs, adjectives and adverbs, these were tagged by combining lexicon- and rule-based methods, whereas all the remaining words were tagged as nouns by default. Finally, all words underwent manual disambiguation and correction.

Our tagging approach coincides to a great extent with that of the new corpus developed concurrently by Kote et al. (2019). There are however a few discrepancies:

1. Verbs expressing modality such as *mund* (can), *duhet* (must), *dua/do* (want) are treated differently. Kote et al. (2019) adopt the `AUX` tag, thereby grouping them together with semantically similar verbs in many other languages including English. We instead use the `VERB` tag, because we want to maintain uniformity in the syntactic annotation of verbal structures introduced with the particle *të* (verbal forms in the subjunctive mood), to be discussed in more detail in Section 5.5.
2. The mediopassive clitic *u*, appearing in the analytical formation of the mediopassive past tense and the gerund form (in which case it is preceded by the particle *duke*) is tagged `PART` by Kote et al. (2019) and `AUX` by us. Our choice is motivated by the fact that it is associated here with the category of voice and/or tense, and that the tag `AUX` in UD v2 is not restricted to verbal auxiliaries.
3. The copula verb *jam* (to be) is tagged `VERB` by Kote et al. (2019) and `AUX` by us, while both corpora use `AUX` when *jam* is used as a temporal auxiliary. The discrepancy here seems to be due to a difference between UD v1, where copula verbs were tagged `VERB`, and UD v2, where `AUX` is the prescribed tag.

In addition to these systematic differences, we have observed that the corpus of Kote et al. (2019) shows some variation in the tagging of ambiguous word forms. For example, the word *të* (to/of) appears both with `PART` and `DET` when occurring in verb groups, while our treebank only uses `PART` in this position.[10] Similarly, the word *që* (that) appears with `CCONJ`, `SCONJ` and `PRON` when introducing relative clauses, while our treebank only uses `PRON` in this position. Most of these differences should be relatively easy to harmonize.

4.5 Syntactic Annotation

The syntactic annotation was performed manually using the annotation tool UD Annotatrix (Tyers et al., 2017), a browser-based tool customized for manual annotation of dependency trees in UD. Applying the UD guidelines to Albanian turned out to be relatively straightforward for the majority of syntactic constructions. In the next section, we discuss some phenomena that gave rise to questions that may be of more general interest to the community.

5 Challenging Constructions

5.1 Core Arguments

One of the fundamental questions when annotating a new language in UD is to determine criteria for distinguishing core arguments from oblique modifiers, including deciding whether there are more than two core arguments. In Albanian, subjects and objects are marked by nominative and accusative case, respectively. In addition, the verb agrees with the subject in person and number; the subject is usually dropped if it is a pronoun. In addition to subjects (`nsubj`) and direct objects (`obj`), we also recognize as core arguments indirect objects (`iobj`) marked by dative case. The argument for treating dative verbal dependents as core is that they behave like (accusative) objects in two important respects. First, they trigger clitic doubling, as discussed in Section 5.2. Secondly, they can permute freely with direct objects when occurring after the verb. Example (6) shows a sentence with three core arguments.

[10] Both corpora consistently use `DET` when it occurs in noun phrases.

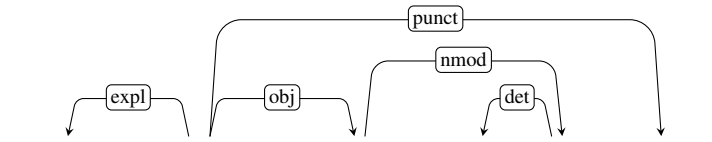
(6) Anna i-a dërgoi Mariës letrën
Anna.NOM her.DAT.CL-it.ACC.CL sent Maria.DAT letter.DEF.ACC
'Anna sent Maria the letter'

5.2 Clitic Doubling

Clitic doubling, the phenomenon where pronominal clitics appear in a sentence along with the noun phrase they refer to – despite being complementary – is widespread in Albanian. It can occur with objects either in the dative or the accusative case. However, while clitic doubling is obligatory for dative objects, it is variable with accusative objects, depending on the focal and topical aspect of the sentence.[11]

In the current treebank, there are two different cases of clitic doubling which are illustrated in examples (7) and (8). In (7), the clitic *E*, which is positioned before the verb, is present along with the nominal it refers to, *qenësinë*. Its presence implies that the focus of the sentence does not lie in the NP itself, otherwise focality would require absence of clitic doubling for the accusative in Albanian (Kapia, 2012). The dedicated solution that UD has for clitic doubling is to treat the full nominal as a core argument and attach the clitic to the verb with the relation `expl` (expletive).

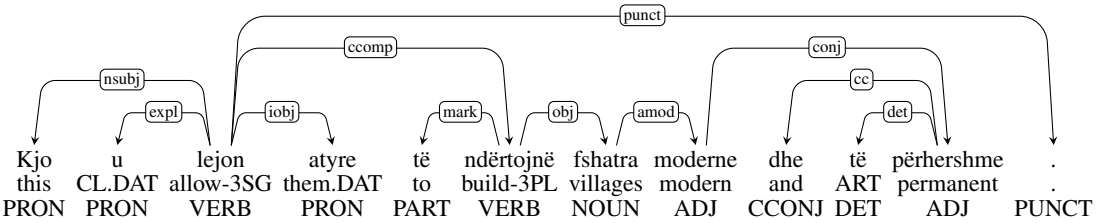
(7)



'They deny her existence.'

Similarly, example (8) features clitic doubling with a dative object, with which clitic doubling always occurs independently of the information structure. Here the same annotation principle was followed, where the nominal is treated as a core argument (here `iobj`), whereas the clitic *u* is an expletive (`expl`). This scheme generally applies when a lexical nominal and its pronominal copy appear together in a sentence.

(8)



'This allows them to build modern and permanent villages.'

It is worth noting here that, although UD does not treat such clitics (co-appearing with their lexical nominal) as arguments, the dative clitics in Albanian (here *u*) appearing with dative objects are grammatically indispensable as opposed to the nominal (here *atyre*). For example, while the lexical nominal *atyre* could be elided with no grammatical or semantic consequences, as in example (9), the same is not true of *u*, which cannot be omitted without loss of grammaticality, as shown in (10).

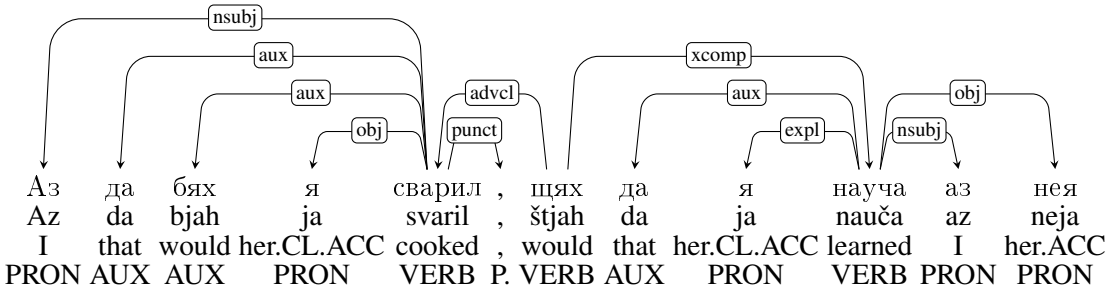
(9) Kjo u lejon të...
this CL.DAT allows to...
'This allows them to...'

(10) *Kjo lejon atyre të...
this allows them.DAT to...
'This allows them to...'

[11] Kapia (2012) and Kallulli (2008) provide more information about the specifics of clitic doubling in Albanian, how focality and topichood affect the presence of a clitic, and how definiteness is also a requirement for clitic doubling in the accusative.

The fact that the clitic is obligatory while the noun phrase or full pronoun is optional might suggest treating the clitic as the core argument and instead use the relation *dislocated* for the co-referential nominal. However, according to our interpretation of the UD guidelines, the dislocation analysis should be used only when the co-occurrence of the pronoun and the nominal is optional. In Albanian though, the full nominal cannot occur without the pronominal clitic (at least in the dative) and the latter is tantamount to an agreement inflection on the verb. The possible omission of the full nominal can therefore be regarded as equivalent to pro-drop. Nevertheless, since the clitic is assigned a syntactic relation (unlike a morphological agreement inflection), we might consider changing the annotation from `expl` to `obj`/`iobj` when the full nominal is omitted. This is in fact what the UD guidelines recommend[12] and what is currently done also in Bulgarian, another language of the Balkan Sprachbund, as shown in example (11):

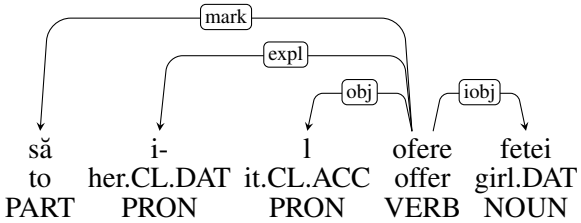
(11)



'If I had cooked it, I would have learned it.'

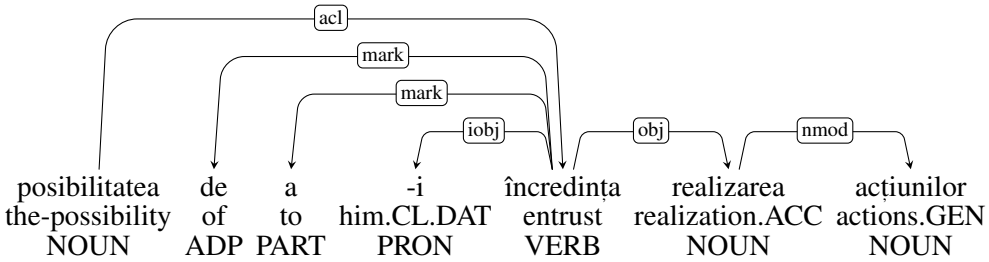
Similarly, a dative clitic is also analyzed as expletive in Romanian when accompanied with a full nominal (12), while it becomes a core argument when the full nominal is missing (13).

(12)



'to give it to the girl'

(13)



'the possibility to entrust him with carrying out the actions'

5.3 Genitive Case-Marking

The genitive case is special in Albanian as its formation requires the use of an article/case marker before the noun[13] with which it forms a constituent. Other languages form the genitive seemingly in the same way: compare the Greek example (14) with Albanian (15).

(14) το βιβλίο της κοπέλας
to vivlío tis kopélas
the.ART.NOM.N book.NOM.N of.ART.GEN.F girl.GEN.F
'the girl's book'

[12] `https://universaldependencies.org/u/dep/expl.html`

[13] Breu (2010) states that the purpose of this article, or "linking particle", as he refers to it, is to morphologically disambiguate the genitive from the dative due to syncretism.

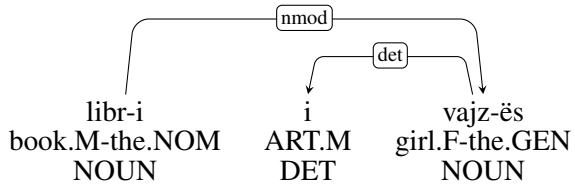
(15) libr-i i vajz-ës
book.M-the.NOM of.ART.M girl.F-the.GEN
'the girl's book'

However, the difference lies in the fact that, in Albanian, the article, here *i*, although syntactically dependent on the possessor, agrees in gender, definiteness and number with the possessum (which is the preceding noun) (Catasso, 2011). For instance, in the example *libri i vajzës*, *i* is masculine, just like the possessum *libri*, despite the possessor *vajzës* being feminine. Nevertheless, the clitic is syntactically dependent on the possessor noun, as is clearly seen in predicative uses of the genitive, where the clitic is separated from the possessum but not from the possessor (Çanta, 2017).

(16) libr-i është i vajz-ës
book.M-the.NOM is ART.M girl.F-the.GEN
'the book is the girl's'

Another question is what kind of modifier the clitic is. One might want to treat it as a case marker, in analogy with the English preposition *of*. However, Albanian *i/e* is considered an article by most linguists; it is also marked for gender, number and case, which aligns with determiners in other Indo-European languages, such as Greek του *(tou)* or German *des*. Therefore, our syntactic annotation treats the clitic as a determiner of the possessor noun, as shown below.

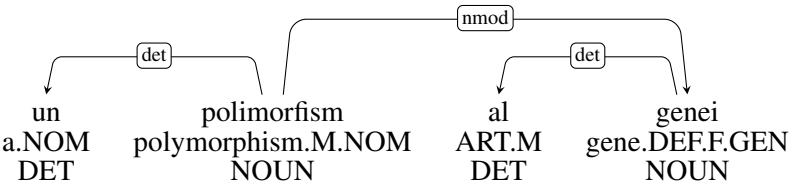
(17)



'the girl's book'

Example (18) shows that a similar phenomenon in genitive constructions can be found in Romanian, where the so-called possessive article also agrees with the possessum rather than the possessor. We therefore believe that the construction should be annotated in the same way in both languages.

(18)



'a polymorphism of the gene'

Finally, it is worth noting that some researchers claim that this multifunctional and controversial article in Albanian have features in common with ezafe (the linking particle) in Persian and other languages (Franco et al., 2015).

5.4 Prearticulation

The peculiar article/particle/clitic *i/e*, except for its presence in the formation of the genitive, is also present in other expressions such as the days of the week (*e martë*, Tuesday) and nominalized adjectives (*të vdekurit*, the dead; *të* is the plural form of *i/e*). However, its most systematic use is in the formation of prearticulated adjectives[14] and pronouns, although there are plain ones as well.

We considered several possible ways of analyzing prearticulated adjectives. One option is to treat them as single words with spaces, but the UD guidelines recommend using this option very restrictively and it would make word segmentation more challenging. Another option is to analyze them as compounds,

[14]Hendriks (1982) does not agree with this term due to the primary function of articles to connect, which according to him is not fulfilled in the case of these adjectives, as their role is contained in the formation of the adjective itself. Therefore, he prefers to refer to them as particle-adjectives.

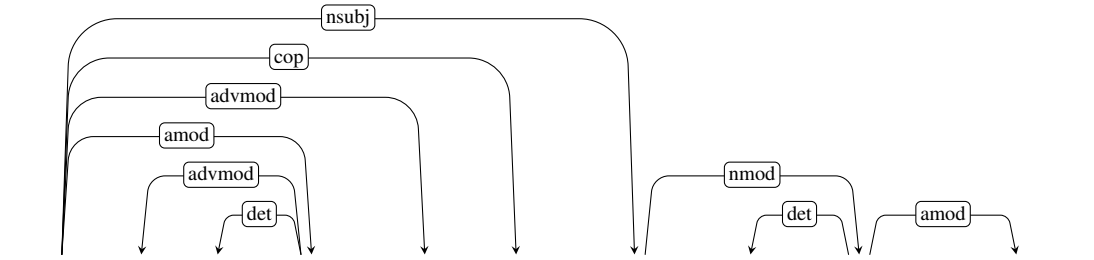
using the `compound` relation, but this was rejected based on the observation that the article is inflected for gender, number and case along with the main adjective and does not add to its meaning, but rather displays an established grammatical phenomenon. Finally, although it is not common for UD to have a determiner depend on a nominal other than a noun (except for cases of ellipsis), the label `det` seemed to us as the most suitable solution for the annotation of such constructions in Standard Albanian.

Nevertheless, even though this annotation is identical to that deployed for the genitive, it should be noted that these are different constructions since prearticulated adjectives are preceded by that article/particle in all cases and the omission of it makes them either ungrammatical or in certain cases leads to changes in part of speech (e.g., *i mirë* good vs. *mirë* well), as opposed to nouns that are preceded by this article only in genitive (e.g., *mendime*, thoughts vs. *i/e mendimeve*, of the thoughts). Examples (19) and (20) illustrate how the noun *mendimeve* appears prearticulated only when in genitive, as opposed to its modifying adjective *të*[15] *bukura*, which is prearticulated by default, independently of the case.

(19) mendime të bukura
thoughts.NOM.F.PL ART.PL beautiful.NOM.F.PL
'beautiful thoughts'

(20) kultivimi i mendimeve të bukura
cultivation.NOM.M.PL ART.M thoughts.GEN.F.PL ART.PL beautiful.GEN.F.PL
'the cultivation of beautiful thoughts'

(21)



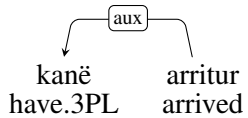
'The most apparent thing at that time was the development of metaphysical thoughts'

As seen in example (21), there are two types of adjectives, *e dukshme* (apparent), which is a prearticulated adjective, and *metafizike* (metaphysical), which is a plain one. There is also a genitive construction present, *i mendimeve* (of the thoughts) which as seen, functions as a nominal modifier to another noun, while itself it is modified by an adjective.

5.5 Modal Verbs

Common auxiliary verbs in Albanian are the copula *jam* (to be) and the verb *kam* (to have), which is used in the formation of the perfect aspect, e.g., *kam shkruajtur* (I have written). These have been assigned the part-of-speech tag `AUX` and the dependency label `aux` when acting as temporal auxiliaries, as in (22). When used as a main verb, *kam* has instead been tagged as `VERB`.

(22)



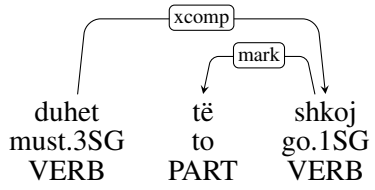
'they have arrived'

However, verbs expressing modality, such as *mund* (can), *duhet* (must) and *do* (will/want), the uninflected form of *dua* (want), have on the contrary been tagged as `VERB`, despite being semantically equivalent to

[15]Here, *të* is the declined form of *i* or *e*. Therefore, in masculine nominative singular, the adjective would be *i bukur*.

modal auxiliaries in some other languages.[16] This treatment is motivated by the subjunctive construction with the particle *të* that the modal verbs govern, e.g., *mund të shkoj* (I may go), *duhet të shkoj* (I must go), and *do të shkoj* (I will go). Consequently, in such constructions the verb in the subjunctive mood depends on the modal verb with the relation `xcomp`, while taking *të* as a *mark* dependent, as shown in (23). A smilar analysis of modal verbs is employed in UD for, e.g., the Slavic languages. There is even a parallelism with the analysis of English *ought*, which combines with a *to*-infinitive, and is analyzed with the same syntactic structure that we propose for the Albanian modal verbs.

(23)

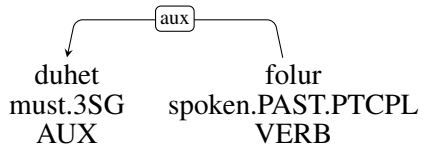


'I must go'

This analysis is parallel to constructions of modal-like verbs with verbal complements, e.g., *shpresoj të kthehem* (I hope to return) and therefore ensures a uniform analysis for all subjunctive constructions introduced by *të*. However, we note that similar constructions are not annotated consistently in all UD treebanks. For example, in Modern Greek (Prokopidis and Papageorgiou, 2017), the analysis of a construction with πρέπει *prépei* (must), which also takes the form of a subjunctive with a particle (να *na*), treats the second verb as the head and assigns the relation `aux` to both the modal verb and the particle.

On the other hand, a drawback of the analysis that we propose for Albanian is that it calls for a different treatment of *duhet* in impersonal constructions, where *duhet* takes a past participle instead of a verb in subjunctive as a complement.[17] An example of this is illustrated in (24).

(24)



'it must be spoken'

The reason behind this different treatment of *duhet* lies in the consistent analysis we aimed to maintain across VP constructions built with a past participle, as in Example (22).

6 Conclusion

Albanian is a morphologically rich language with several grammatical particularities which can prove challenging when trying to find analogies to other languages. In this paper, we presented the first UD treebank for Standard Albanian, which features some of the most characteristic constructions of Albanian. We gave an overview of the formal aspects of the language and analyzed in more detail a few dependency structures that are rather rare or even unique in UD and call for special solutions. Although its current size is not sufficient for the training of tools such as parsers, which in turn could be used for the development of more sophisticated NLP applications, we envision that this starter treebank will encourage further work in the area and will be enlarged in the future.

References

Xhafer Beqiraj. 2014. *Problems of gender accord of two-gender noun determiners*. GRIN Verlag.

[16] Breu (2010) refers to these verbs in Albanian as semi-auxiliary verbs and provides a detailed analysis of all the modals and their usage in this language.

[17] The modal *duhet* (3SG.PRES) is technically the mediopassive form of *dua* (want) and as a regular verb means *he/she/it is needed/wanted*. However, when bearing the modal nuance of obligation, it is always used in the 3rd person singular present tense regardless of the subject. It assumes the meaning of *must* and governs either subjunctive verbs, in the active voice, or past participles, in impersonal constructions.

Walter Breu. 2010. Mood in Albanian. In *Mood in the Languages of Europe*.

Agnesa Çanta. 2017. The category of case in English and Albanian nominal system: A contrastive analysis. *International Journal of English Linguistics*, 7:226, 01.

Nicholas Catasso. 2011. Genitive-dative syncretism in the Balkan sprachbund: An invitation to discussion. *SKASE Journal of Theoretical Linguistics*, 8:70–93, 01.

Ludovico Franco, M. Rita Manzini, and Leonardo M. Savoia. 2015. Linkers and agreement. *The Linguistic Review*, 32(2):277–332.

Besmir Hasanaj. 2009. *A Part of Speech Tagging Model for Albanian*. Lambert Academic Publishing, Saarbrücken, Germany.

Peter Hendriks. 1982. On distinguishing articles in Albanian. *Studies in Slavic and General Linguistics*, 2:95–108.

Klesti Hoxha and Artur Baxhaku. 2019. Albanian language identification in text documents. *CoRR*, abs/1901.04216.

Besim Kabashi and Thomas Proisl. 2016. A proposal for a part-of-speech tagset for the Albanian language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4305–4310, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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, May. 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.

Dalina Kallulli. 2008. Clitic doubling, agreement and information structure: The case of Albanian. In *Clitic Doubling in the Balkan Languages*, page 227–255.

Enkeleida Kapia. 2012. Clitic doubling and information structure in Albanian. *Linguistics*, 50(5):901 – 927.

Nikitas N. Karanikolas. 2009. Bootstrapping the albanian information retrieval. *2009 Fourth Balkan Conference in Informatics*, pages 231–235.

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.

Jouko Lindstedt. 2000. Linguistic balkanization: Contact-induced change by mutual reinforcement. *Studies in Slavic and General Linguistics*, 28, 01.

Ailsa Meechan-Maddon and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120, Paris, France, August. Association for Computational Linguistics.

Maria Morozova and Alexander Rusakov. 2014. Albanian national corpus: Composition, text processing and corpus-oriented grammar development. *Akten der 5. Deutsch-albanischen kulturwissenschaftlichen Tagung*, pages 270–304.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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 International Conference on Language Resources and Evaluation (LREC'20)*. European Language Resources Association (ELRA).

Odile Piton and Klara Lagji. 2008. Morphological study of Albanian words, and processing with NooJ. In *Proceedings of the 2007 International NooJ Conference*, pages 189–205. Cambridge Scholars Publishing.

Prokopis Prokopidis and Harris Papageorgiou. 2017. Universal dependencies for greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106.

Jochen Trommer and Dalina Kallulli. 2004. A morphological analyzer for standard Albanian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. UD Annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17, Prague, Czech Republic.