A Proposal for a Part-of-Speech Tagset for the Albanian Language

Besim Kabashi, Thomas Proisl
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstraße 6, 91054 Erlangen, Germany
besim.kabashi@fau.de, thomas.proisl@fau.de

Abstract

Part-of-speech tagging is a basic step in Natural Language Processing that is often essential. Labeling the word forms of a text with fine-grained word-class information adds new value to it and can be a prerequisite for downstream processes like a dependency parser. Corpus linguists and lexicographers also benefit greatly from the improved search options that are available with tagged data.

The Albanian language has some properties that pose difficulties for the creation of a part-of-speech tagset. In this paper, we discuss those difficulties and present a proposal for a part-of-speech tagset that can adequately represent the underlying linguistic phenomena.

Keywords: Albanian; Tagset; Morphosyntactic annotation

1. Introduction

While there is no specification from EAGLE for Albanian, there are some works about morphological annotation and tagging for the Albanian language, cf. Trommer and Kallulli (2004), Kabashi (2004), Piton et al. (2007), Piton and Lagji (2008), Hasanaj (2012), Kadriu (2013) and Kabashi (2015). In the field of computational morphology, Trommer and Kallulli's (2004) analyzer, which can be used as a kind of tagger, seems to cover the main Albanian inflection types. Its output format follows the EAGLE guidelines standard (Leech and Wilson, 1999), i. e. tags consist of sets of attribute-value pairs, and looks either like this: `[cat:n case:nom num:sg def:+ gen: fem]` or, more compact, like this: `[n nom sg +def fem]`. Alternatively, tags can be collapsed, e. g. the two tags tags `[n nom sg -def fem]` and `[n acc sg -def fem]` can be merged into `[n nom,acc sg -def fem]`. The tagset distinguishes between 17 broad word-class labels, seven of which are reserved for subtypes of pronouns, one for the preposed article, one for sentence equivalents, one for the participle form of the verb, and the rest for the traditional parts of speech, namely nouns, verbs, adjectives, adverbs, particles, prepositions, and conjunctions.

The goal of the tool by Piton, Lagji and Përnaska (2007; 2008), is to cover the inflection of Albanian. Their analysis also allows for preposed articles, if the word forms can have them, cf. the analysis of the adjective *i artë*, engl. *golden*: `i artë,A+FLX=Adj2+ei+m+s`. It is not clear how many tags the tagset has or on how many major word-classes it is based, though it seems to cover at least the ten traditional parts of speech.

The part-of-speech tagging modell for Albanian presented by Hasanaj (2012) consists of a basic tagset with 16 tags and a large tagset with 326 tags. In the basic tagset, there are three tags for delimiters, two for special cases (short forms of pronouns), one for articles, and ten for the traditional parts of speech. The large tagset encodes the word-classes, e. g. `JJ` [=Adj.], `NN` [=Noun], `VB` [=Verb] or `PR` [=Pronoun], and additional features, e. g. Number (Sg. and Pl.). A tag from the large tagset could look like this: `PRDFSE`, which means `PR.` `dem`[onstrative] `fem`[inine] `pl`[ural] `nom`[inative].

Kadriu (2013) uses a tagset of 22 tags. Her tagger covers the morphological categories of orthographic word forms (delimited by spaces or punctuation). In contrast to the traditional ten parts of speech, she distinguishes between feminine and masculine nouns, impersonal, reflexive and transitive verbs, personal and possesive pronouns, determiners, excl[amation], indecl[inable] and indef[inite] elements.

Kabashi's (2015) system extends the traditional parts of speech with additional tags, like abbreviation, e. g. *d.m.th*, or punctuation, e. g. *?* (question mark). For some word-classes, e. g. pronouns, more fine-grained subtypes are specified and have their own tags, e. g. `neve: neve+PersPron+1P+Pl+Dat`. The system can also handle the preposed articles or particles that can occur with some word-classes. Some adverbs, for example, can be preceded by the particle *së*. Depending on whether the particle is present or not, the adverb *pari*, engl. *firstly/in the first place*, for example, is either analysed as `pari: pari+Adv` or as `së pari: së pari+Adv+së`.[1]

So, apart from Piton et al. (2007), Piton and Lagji (2008) and Kabashi (2015), the other tools only cover the word forms without consideration of their articles or particles, i. e. only single graphical tokens in the sense of Cloeren (1999) and Grefenstette (1999).

Word order in Albanian is fairly free, similar to German, which means that the same word forms occuring in different positions in the sentence will often have different grammatical and/or semantic roles. These phenomena cannot be covered by a morphological component that processes word forms in isolation, disregarding their contexts and specifically their articles or particles. Especially the highly frequent "small words" like articles, particles, prepositions etc. have mostly more then one function, some of them even more than ten, e. g. *të*.

It is necessary to take the context into consideration, i. e. (1) to recognise and then to label the linguistic tokens, which can consist of more than one graphical token, and (2) to account for the position of a token in a sentence, i. e. its syntactic function. There is a need for an Albanian tagset that can adequately represent these morphosyntactic phenomena.

[1] There are, of course, other possible analyses of *pari*, e. g. `pari+S+Fem+...` or `1+Ord+...`.

2. Multi-word units

In the context of part-of-speech tagging of Albanian texts, a big challenge is the treatment of multi-word units, i. e. two or more graphical tokens that together make up a single linguistic token. A frequent phenomenon are multi-word units having a preposed article or particle as their first part. These preposed articles or particles can also lead to word-class changes, e. g. from adverb (*gjatë*, engl. *long*) to adjective (*e gjatë*, engl. *long*).

Before Standard Albanian was regulated in 1972, these combinations with preposed articles or particles were written as single graphical tokens by some authors, cf. for example Kristoforidhi (1904, p. 77)[2] where we can find the form ε-γjάτε̱-α [= romanised: *e-gjatë-a*] that stands for *e gjatë* (indefinite) or *e gjata* (definite). The reasoning behind this way of writing the multi-word units was probably to make them easier to read, i. e. to render words that belong together (linguistic tokens) as single graphical tokens.

In written Standard Albanian, there are a lot of cases of "small words", i. e. articles, particles, determiners etc., that are separated from the words they belong to. There are even some grammatical functions that are realized with the help of separately written particles, e. g. graduation of adjectives: *i madh*, engl. *big*, *më i madh*, engl. *bigger*, *më i madhi*, engl. *the biggest*, cf. section 3.2.

3. The tagset

Traditional grammars, e. g. Leonard Newmark and Prifti (1982), Buchholz and Fiedler (1987) or Demiraj et al. (1995), give ten parts of speech for Albanian: Nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, numerals, particles and interjections. Part-of-speech tagsets for use in automatic tagging are usually much more fine-grained, cf. for example Santorini (1990). In the following sections we will discuss all traditional word-classes and will propose a refined tagset.

3.1. Nouns

Nouns have the following morphological categories:

- *Case* (*nominative*, *accusative*, *dative*, *genitive* and *ablative*): The inflectional suffixes for the dative, genitive and ablative forms are identical. The distinction between them can only be made from context.

- *Definiteness* (*indefinite* and *definite*): Indefinite forms like *(një) djalë*, engl. *(one) boy*, can be distinguished from definite forms like *djali*, engl. *(the) boy*.

- *Gender* (*feminine* and *masculine*): A lot of Albanian nouns change gender in the plural form, cf. Demiraj et al. (1995) and Fiedler (2003). The word *art/arti*, engl. *an art/the art*, for example, is masculine in the singular and feminine in the plural (*arte/artet*, engl. *arts/the arts*). Fiedler (2003) calls them *heterogeneous nouns*. This phenomenon makes checking for congruence difficult without a morphological component.

- *Number* (*singular* and *plural*)

There are a couple of "complex" nouns that consist of either a preceding article combined with a noun, e. g. *e hëna*, engl. *saturday*, as in *e hëna është ditë pushimi*, engl. *saturday is a free day*, or of a preceding article combined with an article-adjective, e. g. *i madhi*, engl. *the bigger one*, as in *i madhi është më i lirë se i vogli.*, engl. *the bigger one is cheaper than the smaller*.

A tagset must be able to differentiate between single-word nouns like *hënë/hëna*, engl. *(a) moon/(the) moon*, and "complex" or multi-word nouns like *e hënë/e hëna*, engl. *a monday/the monday*. Also, nouns derived from adjectives and/or verb participles can have a preposed article, e. g. the noun *i pasuri*, engl. *the rich*, which is derived from the adjective *i pasur*, engl. *the rich*, which is in turn derived from the participle form *pasur* of the verb *kam*, engl. *have*.

Similarly, the participle form *pyetur* of the verb *pyes*, engl. *ask*, when preceded by the article *i*, becomes the adjective *i pyetur* as in *Personi i pyetur foli gjatë*, engl. *The asked person spoke for long*. If the suffix *i* is added to the adjective, it will become the noun *i pyeturi* as in *I pyeturi foli gjatë*, engl. *The asked [person] spoke for long*. If it is preceded by the particle *së* instead of the article *i*, it will become the adverb *së pyeturi* as in *Në fund ai s'mbaroi së pyeturi*, engl. *In the end he did't stop asking*.

Depending on the morphological analysis component or the morphological resource, processing *heterogeneous nouns*, e. g. lemmatising them or generating word forms from lemmata, can be problematic. For this reason, there is an extra tag for them.

In the tagset we do not mark number (singular/plural), gender (masculine/feminine), definiteness (definite/indefinite) or case because the goal is not to include a full morphological analysis of nouns but rather to mark those pieces of information that cannot be retrieved using a morphology component or morphological resource.

#	Tag	Name	Example
1	N	Noun	*hënë*
2	NArt	N. w. prep. art.	(*e*) *hënë*; (*të*) *rinjtë*
3	HgN	Heterog. N.	*art*, *-i* sg. m. vs. *-e*, *-et* pl. f.
4	NE	Name	*Peja*; *Drini*; *Joni*;

Table 1: The proposed noun tags

As an example, consider the sentence *Të rinjtë janë nga Peja.*, engl. *The youths are from Peja.*, which is analysed as *Të* \Art *rinjtë* \NArt *janë* \V *nga* \Prep *Peja* \NE . \Punct

3.2. Adjectives

Apart from a few exceptional cases, adjectives have the same morphological categories as nouns. In addition, they have the following properties:

- *Position in the noun phrase*: In the unmarked case, adjectives follow the noun, but they can also occur before the noun, as preposed adjectives, e. g. *i vetmi shkencëtar* vs. *shkencëtari i vetëm*, engl. *the only scientist*.

[2] Transcribed in Albanian letters and edited 1961 in Tirana by A. Xhuvani.

Some adjectives have a preposed article, others do not. Both need to be congruent to the noun they modify. This is marked in the inflectional suffixes and in the preposed article. There are also a few noninflected adjectives, e. g. *blu*, engl. *blue*.

- *Graduation*: Adjectives in Albanian have three grades, *positive*, *comparative*, and *superlative*. Graduation is realized as a combination of the base word with the comparative particle *më*, e. g. positive *i mirë* engl. *good*, comparative *më i mirë* and superlative *më i miri*. This particle has its own tag, cf. Section 3.10.

We propose five tags for describing adjectives in Albanian. There are separate tags for adjectives that occur before nouns, for adjectives with preposed articles, and for noninflectional adjectives.

#	Tag	Name	Example
5	Adj	Adjective	[*djali*] *trim*
6	PPAdj	Preposed adj.	*trimi* [*djalë*]
7	AdjPPArt	Adj. w. art.	[*djali*] *i mirë*
8	PPAdjPPArt	PPAdj. w. art.	*i miri* [*djalë*]
9	AdjNIf	Noninfl. adj.	*blu/neto*

Table 2: The proposed adjective tags

The sentence *Ata studiojnë artet e bukura në Akademinë e Arteve.*, engl. *They study fine arts at the College of Arts.*, for example, is tagged as *Ata* `\Pron` *studiojnë* `\V` *artet* `\HgN` *e* `\Art` *bukura* `\AdjPPArt` *në* `\Prep` *Akademinë* `\N` *e* `\Art` *Arteve* `\N`. `\Punct` .

3.3. Numerals

Numerals in Albanian are subclassified in cardinal and ordinal numbers. Ordinal numbers have the same properties as adjectives, except graduation, and are always preceded by an article.

#	Tag	Name	Example
10	NumC	Cardinal number	*dy* [fitore]
11	NumO	Ordinal number	[fitorja] *e dytë*

Table 3: The proposed numeral tags

As an example, consider the sentence *Sot ishte fitorja e dytë e tij brenda një jave.*, engl. *This was his second victory within one week.*, that is tagged as *Sot* `\Adv` *ishte* `\V` *fitorja* `\N` *e* `\Art` *dytë* `\NumO` *e* `\Art` *tij* `\PossP` *brenda* `\Prep` *një* `\NumC` *jave* `\N`. `\Punc`.

3.4. Pronouns

Pronouns can be classified into subtypes according to their specificity. An interrogative pronoun or a personal pronoun is different from a relative pronoun.

We have an extra tag for each commonly distinguished type of pronouns. Similar to nouns or adjectives, some pronouns can be preceded by a preposed article. This can, for example, turn the interrogative pronoun *cili*, engl. *who/which* into the relative pronoun *i cili*, engl. *who/which*. For this reason, article and pronoun need to be treated together.

Table 4 lists all proposed tags that decribe pronouns. As can be seen, there are four types of pronouns that are preceded by an article.

#	Tag	Name	Example
12	PersP	Personal pron.	*ti*
13	DemP	Demonstr. pron.	*ky/këta*
14	DemPPPArt	DemPron w. art.	*i tillë*
15	PossP	Possesive pron.	*im*
16	PossPPPArt	PossP w. prep. art.	*i tij/të vetën*
17	IntP	Interrogative pron.	*kush*
18	IntPPPArt	IntP w. art.	*i kujt/i cilit*
19	RelP	Relative pronoun	*që*
20	RelPPPArt	RelP w. art.	*i cili*
21	IndefP	Indefinite pron.	*dikush*
22	ReflP	Reflexive pron.	(*me*) *vete*

Table 4: The proposed pronoun tags

3.5. Verbs

Verbs have the greatest number of grammatical categories, cf. Leonard Newmark and Prifti (1982). They are as follows:

- *person* (*1st*, *2nd*, *3rd*)
- *number* (*singular* and *plural*)
- *voice* (*active* and *non-active*, i. e. *passive*, *middle*, *reflexive* or *reciprocal*)
- *mood* (*indicative*, *subjunctive*, *optative*, *admirative* and *imperative*)
- *tense* (*present*, *past* and *future*)
- *aspect* (*common*, *perfect*, *progresive*, *inchorative*, *definite* and *imperfect*)
- *finiteness* (*finite* and *non-finite*, i. e. *infinitive*, *participle*, *gerundive* and *absolutive*)

We will not go into further detail here and will only briefly discuss some properties that are important in the context of constructing a tagset, namely those forms that combine with particles like *të* or *do*, cf. also Section 3.10.

- In the passive voice, in some cases, verbs are preceded by the article *u*. This must be treated as an extra form, having its own tag(s).
- In Albanian, the verbal complex can also include pronominal clitics. The passive particle can be combined with these pronominal clitics, e. g. *u* + *e* → *ua*. This must also be treated as an extra form, with separate tag(s). Some of these combinations are written with an apostrophe, e. g. *m'u*.
- Some constructions, e. g. subjunctive mode, or the future tense, have additional particles, e. g. *të* and *do*. These particles are listed in Section 3.10.

Here are some examples:

> In the sentence *Libri u botua.*, engl. *The book is published.*, the verbal complex will be tagged *u*\PassP *botua*\VpP.
>
> In the sentence *Ai do ta blejë librin.*, engl. *He will bay the book.*, the verbal complex will be tagged *do*\FutP *ta blejë*\VSubjPPCl.
>
> In the sentence *Ai nuk do të na i kthejë neve librat sot.*, engl. *He will not give back us the books today.*, the verbal complex including the negation particle will be tagged *nuk*\NegP *do*\FutP *të*\SubjP *na*\PCl *i*\PCl2 *kthejë*\VSubjPPCl.

Furthemore we differentiate between auxiliar, modal, reflexive and reciprocal verbs as well as the participle form.

#	Tag	Name	Example
23	V	Verb	*tha*
24	VP	Verb	*thënë*
25	VSPPC	V. w. SPart and cl.	*ta tha*
26	VpP	V. w. passive particle *u*	*u tha*
27	VPC	V. w. cl.	*i tha*
28	VpPC	V. w. pass. part. and cl.	*ua tha*
29	VAux	Auxiliar verb	*kam*
30	VMod	Modal verb	*mund*
31	VRefl	Reflexive verb	*lahem*
32	VRecp	Reciprocal verb	*përshëndeten*

Table 5: The proposed verb tags

3.6. Adverbs

Some adverbs, like members of other word-classes mentioned above, can occur with a preceding article, e. g. *së voni*, engl. *last/late*. There are also adverbs that consist of multiple graphical tokens, e. g. *kohë pas kohe*, engl. *from time to time*. All the individual graphical tokens of such multi-word adverbs will be tagged as parts of a multi-part adverb, e. g. *kohë*\AdvMP *pas*\AdvMP *kohe*\AdvMP. The graphical tokens that make up the adverb are adjacent, therefore this approach is sufficient to infer that a sequence of AdvMP tags consitutes a single multi-word adverb.

#	Tag	Name	Example
33	Adv	Adverb	*mirë*
34	AdvA	Adverb w. prep. part.	*së shpejti*
35	AdvMP	Multi-token Adverb	*kohë pas kohe*

Table 6: The proposed adverb tags

3.7. Conjunctions

In addition to coordinating and subordinating conjunctions, there are some two-part conjunctions where the two parts are in a non-contact position in the sentence, e. g. *aq … sa* as in *aq mirë sa ne u befasuam*, engl. *so good that we were suprised.* In order to be able to identify those tokens as being parts of a multi-word unit, we introduce separate tags for the first and the second part of such two-part conjunctions.

#	Tag	Name	Example
36	CConj	Coordinating conj.	*dhe*
37	SConj	Subordinating conj.	*që*
38	ConjP1	Conj. part one	*edhe*$_{[P1]}$… *edhe* …
39	ConjP2	Conj. part two	*edhe* … *edhe*$_{[P2]}$…

Table 7: The proposed conjunction tags

3.8. Prepositions

The members of this word-class could be subcategorized according to their grammatical and semantic roles, potentially benefitting downstream processes like parsers. However, we decided not to make any further distinctions as these are not always clear-cut.

#	Tag	Name	Example
40	Prep	Preposition	*me*; *pa*; *nga*; *për*;

Table 8: The proposed preposition tag

3.9. Interjections

Interjections are not inflected in Albanian. Since it is a closed word-class, its members can be listed in a lexicon.

#	Tag	Name	Example
41	Intj	Interjection	*o*; *hm*; *uh*; *ii*;

Table 9: The proposed interjection tag

3.10. Particles

Some particles influence not only one or two words (local influence), but the whole sentence (global or sentence-level influence), e. g. negation particles like *nuk*, engl. *not*. The negation particle *s'* occurs in amalgamated form with the word which it belongs to, e. g. *s'punon*, engl. *it doesn't work* as opposed to *punon*, engl. *it works*.

Negation in Albanian can be realized as double negation, as in *nuk fiton dot* or *s'fiton dot*, engl. *he can not win (in any way)*. Compare this to simple negation: *nuk fiton* or *s'fiton*, engl. *he can not win*. In contrast to words like *assesi*, engl. *no way*, which are adverbs and can replace *dot*, we propose to tag the latter as *NegPartD* because it is a semantically unspecified general negation in the role of double negation. In cases where *dot* does not act as double negation marker, e. g. in *E bën dot këtë punë?*, engl. *Can you really (= dot) do this job?*, it will be tagged as an adverb.

The particle *do* has a separate status because it occurs in the verbal complex for building a future tense. We also propose a separate tag for the particle *po* in its function as aspect marker, e. g. *ai po punon*, engl. *he is working* vs. *ai punon*, engl. *he works*.[3]

[3] *Po* cannot only be an aspect marker but has several meanings and functions, e. g. *po aq*, engl. *same*.

#	Tag	Name	Example
42	Part	Particle	*ja*
43	AdjComPart	Comp. part.	*më* [*i mirë*]
44	FutP	Future part.	*do*
45	NegP	Neg. part.	*nuk/mos/jo*
46	ProhP	Prohib. part.	*mos*
47	NegPartS	Neg. part. *s'*	*s'*[*punon*]
48	NegPartD	Neg. part. *dot*	[*s'/nuk* . . .] *dot*
49	SubjP	Subj. part. *të*	*të*
50	JusP	Juss. part. *le*	*le*
51	InfP	Infin. part. *për*	*për*
52	GerP	Gerund. part. *duke*	*duke* [*punuar*]
53	PartPa	Part. *pa*	*pa* [*punuar*]
54	QPA	Question part. A/a	*A* [*punon*]*?*
55	CondP	Cond. part. *në/po*	*në/po* . . .
56	PartPCl	Particle and clitic	*m'u* [i. e. *më*+*u*]
57	Asp	Progress. part. *po*	*po* [*lexon*]

Table 10: The proposed particle tags

The "small words" are a challenge for the tagging process, because there is a lot of overlap and ambiguity both within the closed classes and between some function words and word forms from other word-classes. Here are some examples for such syncretisms:
The word-form *do* is either a particle (*FutP*) or a form of the verb *dua*, engl. *to wish* (*V*); *e* is either a conjunction (*CConj*) or an article (*Art*).
The particle *mos* can occur in the verbal complex as a negation particle (*NegP*), e. g. *për të mos thënë*, engl. *not to say*, or it can be a prohibitive particle (*ProhP*) as in *Mos ecni shpejt!*, engl. *Do not walk so fast!*
The word-form *duke* can occur in the verbal complex as a gerundive particle (*GerP*), e. g. *duke punuar*, engl. *(while) working*, or it can be a finite passive form of the verb *dukem* (*VpP*) as in *U duke bukur!*, engl. *You looked beautiful!*
The particle *për* can occur in the verbal complex, with the subjunctive particle *të* (*SubjP*) to constract an infinitive of an verb, e. g. *për të thënë*, engl. *to say* as in *Ne kemi për të thënë diçka.*, engl. *We will say something.*
The particle *le* can either occur with the subjunctive particle *të* (*SubjP*) to construct the jussive mood of a verb, e. g. *le të vrapojmë*, engl. *let's walk. Le* can also occur as a finite form of the verb *lë* (*V*) as in *Ti le librat atje.* , engl. *You have left the books there.*

3.11. Articles and clitics

Preposed articles have their own tag, *Art*. This tag is used for all preposed articles of nouns, adjectives and pronouns. Similar to two-part conjunctions there are also pairs of clitics. However, the clitics that can make up the first part of a pair can also occur on their own. Therefore, we have separate tags for clitics that occur on their own or as the first part of a pair (*PCl*, e. g. *Ai i*\PCl *ktheu librat.*, engl. *He gave back the books.*) and clitics that occur as the second part of a pair (*PClP2*, e. g. *Ai na*\PCl *i*\PCl2 *ktheu neve librat.*, engl. *He gave us back the books.*).
The tag *SubjPPCl* is used for words that are a blend of a subjunctive particle and a pronominal clitic, e. g. the word *te* that is a blend of *të* and *e*. As an example, consider the sentence *Ai do ta lexojë librin.*, engl. *He will read the book.*. Here, the verbal complex is tagged *do*\FutP *ta*\SubjPPCl *lexojë*\VSubjPPCl.

#	Tag	Name	Example
58	Art	Article	*i/e/të/së*
59	PCl	Pronominal clitic	*i*
60	PClP2	Second part of pron. cl.	*e* [in: *na e*]
61	SubjPPCl	Subjunc. part. and PCl	*ta* [i. e. *të*+*e*]

Table 11: The proposed article and clitic tags

3.12. Abbreviations

All abbreviations are tagged with the tag *Abbr*.

#	Tag	Name	Example
62	Abbr	Abbreviation	*d.m.th.*; *etj.*; *km.*; *TVSH*

Table 12: The proposed abbreviation tag

3.13. Foreign words

Foreign words are to be marked with the symbol *FW*. Usually, these are words which cannot be recognized by a morphology component or looked up in a lexical or morphological resource.

#	Tag	Name	Example
63	FW	Foreign word / Non-Alb.	*web*

Table 13: The proposed tag for foreign words

3.14. Punctuation

We distinguish between two groups of punctuation marks: (1) potentially sentence-ending punctuation like periods or exclamation marks, and (2) other punctuation.

#	Tag	Name	Example
64	Punct	Punctuation basic	. *? !*
65	Punct2	Punctuation extended	, : ; - – . . .

Table 14: The proposed punctuation tags

3.15. Non-linguistic elements

Non-linguistic elements like *§* or % also have their own tag. Usually, these are various kinds of symbols like *A*+.

#	Tag	Name	Example
66	NLE	Non-linguistic element	· § % . . .

Table 15: The proposed tag for non-linguistic elements

3.16. Emoticons

Emoticons are used in a wide variety of texts, e. g. e-mails or chats, to represent, for example, laughter, irony, or other emotions. We put all of them in one group and tag them as *EM*. The primary goal is to separate those entities from the rest of text.

#	Tag	Name	Example
67	EM	Emoticon	*:-)*

Table 16: The proposed emoticon tag

4. Evaluation corpus

We are currently in the process of manually annotating a sample of approximately 2 000 sentences. Half of those sentences have been selected randomly from texts of different genres, the other half has been selected manually in order to allow for a wider variety of linguistic phenomena in the sample corpus.

Since the corpus has not been completely tagged, yet, the tagset might still be slightly revised or refined during the tagging process. Once the sample corpus is completely annotated, we will perform evaluation experiments with automatic part-of-speech taggers.

5. Conclusions and future work

With regard to the morphological complexity of its inflectional system, Albanian is comparable to German. The tagset presented here consists of 67 tags and aims to adequately represent the morphosyntactic properties of the Albanian language. In particular, combinations of preposed articles or particles with words of other word-classes are treated in a linguistically sensible way. This should also benefit downstream processes like chunkers or parsers.

A number of morphological properties of certain word-classes are not accounted for by the tagset, e. g. gender (nouns) or number (nouns, verbs). Those properties were left out on purpose, (1) because they would have dramatically increased that tagset, and (2) because they are readily available from morphological analyzers like Trommer and Kallulli (2004), Piton and Lagji (2008) or Kabashi (2015).

We would like to stress once more that the evaluation corpus has not been fully annotated, yet, so the tagset presented here might still undergo some minor changes. We would also like to note that this tagset could also be extended to a larger version which further subclassifies adverbs, particles, prepositions and conjunctions.

Once the annotation of the sample corpus is completed, we will leverage automatic part-of-speech taggers and semi-automatically annotate a larger corpus of Albanian texts. That larger corpus might then be used to train reasonably good tagger models to allow for fully automatic tagging.

Morphosyntactic part-of-speech tagging is an important building block of NLP tool chains and we hope that the tagset presented here will be useful in developing tools for more sophisticated analyses like syntactic parsers for Albanian.

6. Bibliographical References

Buchholz, O. and Fiedler, W. (1987). *Albanische Grammatik*. VEB Verlag Enzyklopädie, Leipzig.

Cloeren, J. (1999). Tagsets. In van Halteren (1999), pages 37–54.

Demiraj, Sh., Agalliu, F., Agoni, E., Dhrimo, A., Hysa, E., Lafe, E., and Likaj, E. (1995). *Morfologjia*, volume 1 of *Gramatika e Gjuhes Shqipe*. Akademia e Shkencave e Republikës së Shqipërisë, Tiranë.

Fiedler, W. (2003). Albanisch. In Thorsten Roelcke, editor, *Variationstypologie / Variation Typology*, pages 749–797. de Gruyter.

Grefenstette, G. (1999). Tokenization. In van Halteren (1999), pages 117–133.

van Halteren, H., editor. (1999). *Syntactic Wordclass Tagging*. Kluwer Academic Publishers, Dordrecht.

Hasanaj, B. (2012). *A Part of Speech Tagging Model for Albanian*. Lambert Academic Publishing, Saarbrücken.

Kabashi, B. (2004). Analiza automatike e fjalëformave të gjuhës shqipe. In *Seminari XXIII Nërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*, pages 129–135. Universiti i Prishtinës, Prishtinë.

Kabashi, B. (2015). *Automatische Verarbeitung der Morphologie des Albanischen*. FAU University Press, Erlangen.

Kadriu, A. (2013). NLTK tagger for Albanian using iterative approach. In *Proceedings of the 35th International Conference on Information Technology Interfaces (ITI)*.

Kristoforidhi, K. (1904). Λεξικόν της Αλβανικής Γλώσσης *(Fjalor shqip-greqisht; Albanian-Greek Dictionary)*. Sakellarios, Athens.

Leech, G. and Wilson, A. (1999). Standards for tagsets. In van Halteren (1999), pages 55–80.

Leonard Newmark, P. H. and Prifti, P. (1982). *Standard Albanian – A Reference Grammar for Students*. Stanford University Press, Stanford, CA.

Piton, O. and Lagji, K. (2008). Morphological study of Albanian words, and processing with NooJ. In Xavier Blanco et al., editors, *Proceedings of the 2007 International NooJ Conference*, pages 189–205. Cambridge Scholars Publishing.

Piton, O., Lagji, K., and Përnaska, R. (2007). Electronic dictionaries and transducers for automatic processing of the Albanian language. In Zoubida Kedad, et al., editors, *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007*, pages 407–413. Springer, Berlin, Heidelberg.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project.

Trommer, J. and Kallulli, D. (2004). A morphological analyzer for standard albanian. In *Proceedings of LREC'2004*, pages 1271–1274.