Albanian Part-of-Speech Tagging: Gold Standard and Evaluation

Besim Kabashi, Thomas Proisl
Friedrich-Alexander-Universität Erlangen-Nürnberg
Professur für Korpuslinguistik
Bismarckstr. 6, 91054 Erlangen, Germany
{besim.kabashi, thomas.proisl}@fau.de

Abstract

In this paper, we present a gold standard corpus for Albanian part-of-speech tagging and perform evaluation experiments with different statistical taggers. The corpus consists of more than 31,000 tokens and has been manually annotated with a medium-sized tagset that can adequately represent the syntagmatic aspects of the language. We provide mappings from the full tagset to both the original Google Universal Part-of-Speech Tags and the variant used in the Universal Dependencies project. We perform experiments with different taggers on the full tagset as well as on the coarser tagsets and achieve accuracies of up to 95.10%.

Keywords: Albanian, part-of-speech tagging, gold standard

1. Introduction

Albanian is an Indo-European language that is spoken by approximately 5.4 million people. Within the Indo-European family of languages, Albanian constitutes a subgroup of its own; it is a distinct branch on the same level as, for example, the Hellenic, Romance, Slavic or Germanic subgroups. The language has a diverse vocabulary with many loan words due to language contact with Greek, Latin/Italian, Slavic languages and Turkish.

Albanian has a rich morphological system and a relatively free word order, similar to, for example, German. Particularly challenging, from a pos tagging point of view, are the many multi-word units. An interesting and frequent phenomenon are multi-word units that have articles or particles as their first part. These combinations are borderline cases at the lexis-grammar interface. They are certainly more grammatical in nature and written as two separate graphical tokens. Here is one example from standard Albanian: *i mirë* "(the) good one" (masc.) vs. *e mirë* "(the) good one" (fem.). Presence or absence of the preceding article or particle can change the meaning of a word or its part of speech.

Albanian is one of the Balkan languages with the least resources available. In particular, there is no part-of-speech tagged corpus available,[1] let alone one that covers the multi-word phenomena mentioned above. This paper is a step to improve that situation. In the next section, we briefly discuss previous work in the areas of part-of-speech tagging and morphological analysis. In Section 3., we present a new manually annotated gold standard corpus for part-of-speech tagging. The corpus is annotated with a medium-sized tagset that can adequately represent the multi-word phenomena mentioned above. To improve interoperability with existing multilingual tools and resources, we provide mappings to the popular Universal Part-of-Speech Tags, both the Google and the Universal Dependencies variant (Section 4.). In Section 5., we perform evaluation experiments with different part-of-speech taggers on the full tagset and on the coarser UPOS tagsets.

[1]The Albanian National Corpus (`http://web-corpora.net/AlbanianCorpus/search/`), which consists of roughly 16.7 million tokens, is only annotated with morphological analyses that have not been disambiguated (and only for words known to the morphological analyzer). The AlCo corpus (Kabashi, 2017), which consists of roughly 100 million tokens, is part-of-speech tagged, but is not yet publicly available.

2. Related Work

2.1. Part-of-Speech Tagging

Hasanaj (2012) presents two tagsets: A basic set that consists of 16 tags and a large set that consists of 326 tags. In the basic tagset, there are ten tags for the traditional parts of speech, three tags for delimiters, two for special cases (short forms of pronouns) and one for articles. The large tagset encodes the major word-classes and additional features like number or case. (Hasanaj, 2012) also attempts an evaluation of the two tagsets using a maximum entropy tagger and a perceptron tagger. However, his gold standard corpora consist only of 263 tokens for the basic tagset and 641 tokens for the large tagset.

Kadriu (2013) uses a tagset of 22 tags that refines the ten traditional parts of speech in some places and adds tags for feminine and masculine nouns, impersonal, reflexive, transitive and intransitive verbs, personal and possessive pronouns, determiners, exclamations, indeclinables and indefinite elements. Her tagging system is implemented on top of the NLTK unigram and regular expression taggers and uses a simple stemming algorithm to deal with unknown words. Her evaluation is based on 30 news articles from three domains. The number of tokens in the gold standard corpus is not mentioned in the paper.

The tagsets by Hasanaj (2012) and Kadriu (2013) are either very small or extremely large and do not cover the interactions between words and their preceding articles or particles described in Section 1. In our own previous work (Kabashi and Proisl, 2016), we introduce a medium-sized tagset for Albanian that consists of 67 tags and that 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.

2.2. Morphological Analysis

There are a number of tools for the morphological analysis of Albanian that provide very detailed analyses but no disambiguation and no mapping to a medium-sized tagset.
Trommer and Kallulli (2004) present a morphological analyzer that seems to cover the main inflection types of Albanian. Its output format follows the EAGLE guidelines standard (Leech and Wilson, 1999) in that the tags consist of sets of attribute-value pairs. The tags come in three varieties: Detailed, abbreviated and collapsed. We can gather from the paper that 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. Trommer and Kallulli (2004) evaluate their morphological analyzer against a gold standard corpus consisting of 1,000 tokens.
The goal of the tool described in Piton et al. (2007) and Piton and Lagji (2008) is to cover the inflection of Albanian. For words forms that can have them, their analysis also allows for preposed articles. It is not clear how many morphological tags they have or on how many major word-classes the tagset is based, though it seems to cover at least the ten traditional parts of speech. Their papers do not include an evaluation.
UniParser (Arkhangelskiy et al., 2012), the morphological analyzer used in the Albanian National Corpus, supports a variety of languages, e. g. Albanian, Greek, Kalmyk, Lezgian and Ossetic. There seems to be no published information about the Albanian model.
The morphological analyzer and generator by Kabashi (2015) extends the traditional parts of speech with additional tags for things like abbreviations or punctuation marks. For some word-classes, e. g. pronouns, more fine-grained subtypes are specified and are given their own tags. The system can also handle the preposed articles or particles that can occur with some word-classes. The coverage of the morphological analyzer is evaluated against word lists that each comprise more than 100,000 entries.

3. Gold Standard

3.1. The Tagset

The tagset used to annotate the corpus is a revised version of our earlier draft (Kabashi and Proisl, 2016) that has been used in the corpus described by Kabashi (2017). Traditional grammars like Newmark et al. (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. While we follow this general division, the tagset allows for more fine-grained distinctions within the traditional word-classes and has additional tags for things like interjections, articles, pronominal clitics, punctuation, etc. that do not fit into the traditional word-classes.
In the course of annotating the corpus, we came across phenomena that could not be adequately analyzed with the old version of the tagset. Therefore, we refined and extended it by introducing ten novel tags. We also renamed some of the existing tags to have a more consistent and transparent naming scheme. In particular, we make the following changes:

- We introduce four new verb tags: VImpv (imperative form of verb), VImpvCl (imperative with infixed clitic), VSubj (verb with subjunctive particle) and VSubjPass (verb with subjunctive and passive particles).
- We introduce three new tags for pronouns: PPersSF (short form of personal pronoun), PIndefA (indefinite pronoun preceded by article) and PIntÇ (interrogative pronoun *Ç/ç*).
- For multipart conjunctions, we distinguish between coordinating (ConjC1, ConjC2) and subordinating (ConjS1, ConjS2).
- We introduce two new particle tags: PtMod (modal particle), PtPass (passive particle) and PtQM (question particle *mos*).
- Several tags were renamed to make them more consistent with the rest of the tagset. Now, all tags start with the major word-class (N, V, Adj, etc.) and are easier to interpret, e. g. PPAdjPPArt vs. AdjPA for a preposed adjective with preceding article or RelPPPArt vs. PRelA for a relative pronoun with preceding article.

The complete tagset is shown in Table 1.

3.2. The Corpus

We manually annotated a sample of 2,020 sentences (31,584 tokens) with the part-of-speech tagset described above. Half of the sentences have been randomly selected from a large collection of texts consisting of literary works, news and science articles and web texts. Selection of the other half has been guided by the wish to include as wide a variety of linguistic phenomena as possible in the gold standard corpus. Therefore, these sentences were manually selected and contain instances of rarer morphosyntactic phenomena. Consequently, the resulting corpus is definitely less homogeneous than a collection of whole documents and has a higher type/token-ratio.

The data was annotated by two native speakers who are trained linguists. For creating the gold standard, all controversial cases were discussed with a non-native speaker of the language who is a trained linguist. Table 2 gives agreement scores between the annotators and the gold standard. For this purpose, we also created two additional versions of the corpus where we mapped the manual annotations to two flavors of Universal Part-of-Speech Tags (cf. Section 4.).

The major differences between the annotators are visualized in Fig. 1. For this visualization, we map the tags to the coarser tagset used in the Universal Dependencies project (cf. the next section). As we can see, the two largest areas of disagreement between the annotators are within-class choices for pronouns and verbs. The annotators also disagree fairly often on whether a word is a particle, coordinating or subordinating conjunction. Fig. 2 gives a detailed visualization of the differences within the pronoun class. The major source of disagreement is the choice between PCl, i. e. pronominal clitics, and PClSubj, i. e. amalgamations of subjunctive particle and pronominal clitic. Fig. 3 visualizes the differences within the verb class. Major sources of disagreement are the choices between V and VPass, V and VSubjCl, and VPass and VRefl.

#	Tag	Name	Example
1	**N**	Noun	*hënë*
2	**N**A	Noun preceded by article	*e **hënë***
3	**N**Hg	Het. noun	*art* (sg. m.) vs. *arte* (pl. f.)
4	**N**E	Name	*Peja/Drini/Joni*
5	**V**	Verb (finite forms)	*tha*
6	**V**Part	Participle (non-finite forms)	*thënë*
7	**V**Cl	V. w. clitic	*i **tha***
8	**V**Impv	Imperative form	*Prit!/Fol!*
9	**V**ImpvCl	Imperative w. clitic	*Tregomëni!*
10	**V**Pass	V. w. pass. part. *u*	*u **tha***
11	**V**PassCl	V. w. pass. part. and clitic	*ua **tha***
12	**V**Subj	V. w. subj. particle	*të **thotë***
13	**V**SubjCl	V. w. pass. part. and cl.	*ta **tha***
14	**V**SubjPass	V. w. subj. a. pass. part.	*t'u **thotë***
15	**V**Aux	Auxiliar verb	*kam*
16	**V**Mod	Modal verb	*mund*
17	**V**Recp	Reciprocal verb	*njihen*
18	**V**Refl	Reflexive verb	*lahem*
19	**Adj**	Adjective	*djali **trim***
20	**Adj**A	Adj. preceded by article	*djali i **mirë***
21	**Adj**P	Preposed adj.	***trimi** djalë*
22	**Adj**PA	Prep. adj. prec. by art.	*i **miri** djalë*
23	**Adj**N	Noninflected adjective	*blu/neto*
24	**Adv**	Adverb	*mirë*
25	**Adv**Pt	Adv. prec. by part.	*së shpejti*
26	**Adv**MP	Multipart adverb	*kohë pas kohe*
27	**P**Pers	Personal pronoun	*ti*
28	**P**PersSF	Pers. pron. (short form)	*ta/të/to*
29	**P**Dem	Demonstrative pronoun	*ky/këta*
30	**P**DemA	PDem preceded by article	*i tillë*
31	**P**Poss	Possessive pronoun	*im*
32	**P**PossA	PPoss preceded by article	*i tij/të vetën*
33	**P**Int	Interrogative pronoun	*kush*
34	**P**IntA	PInt preceded by article	*i kujt/i cilit*
35	**P**IntÇ	Interrogative pronoun *Ç/ç*	***ç'**libër?*
36	**P**Rel	Relative pronoun	*që*
37	**P**RelA	PRel preceded by article	*i cili*
38	**P**Indef	Indefinite pronoun	*dikush*
39	**P**IndefA	PIndef preceded by article	*të tjerëve*
40	**P**Refl	Reflexive pronoun	*me **vete***
41	**Prep**	Preposition	*me/pa/nga/për*
42	**Conj**C	Coordinating conjunction	*dhe*
43	**Conj**S	Subordinating conjunction	*që*
44	**Conj**C1	First part of coord. conj.	***edhe** … edhe*
45	**Conj**C2	Second part of coord. conj.	*edhe … **edhe***
46	**Conj**S1	First part of subord. conj.	***le që** … por*
47	**Conj**S2	Second part of subord. conj.	*le që … **por***
48	**Num**C	Cardinal number	***dy** fitore*
49	**Num**O	Ordinal number	*fitorja e **dytë***
50	**Pt**	Particle	*ja*
51	**Pt**Comp	Comparative particle	***më** i mirë*
52	**Pt**Cond	Conditional particle *në/po*	*në/po …*
53	**Pt**Fut	Future particle	*do*
54	**Pt**Ger	Gerundive particle *duke*	***duke** ecur*
55	**Pt**Inf	Infinitive particle *për*	*për*
56	**Pt**Jus	Jussive particle *le*	*le*
57	**Pt**Mod	Modal particle	*mund*
58	**Pt**Neg	Negation particle	*nuk/mos/jo*
59	**Pt**NegD	Negation particle *dot*	*s'/nuk … **dot***
60	**Pt**NegS	Negation particle *s'*	***s'** punon*
61	**Pt**Pass	Passive particle *u*	*u*
62	**Pt**PassCl	Pass. part. with clitic	*iu*
63	**Pt**Priv	Privative particle *pa*	***pa** punuar*
64	**Pt**Prog	Progressive particle *po*	***po** lexon*
65	**Pt**Proh	Prohibitive particle	*mos*
66	**Pt**QA	Question particle *A/a*	***A** punon?*
67	**Pt**QM	Question particle *mos*	***Mos** iku?*
68	**Pt**Subj	Subjunctive particle *të*	*të*
69	**Intj**	Interjection	*o/hm/uh/ii*
70	**Art**	Article	*i/e/të/së*
71	**PCl**	Pronominal clitic	*i*
72	**PCl**2	2nd part of pron. clitic	*e* [in: *na e*]
73	**PCl**Subj	Subj. particle and PCl	*ta* [i. e. *të+e*]
74	**Abbr**	Abbreviation	*d.m.th./etj.*
75	**FW**	Foreign word/Non-Alban.	*web*
76	**Punct**	Punctuation (sent.-ending)	*. ? !*
77	**Punct**2	Punctuation (not sent.-end.)	*, : ; - –*
78	**NLE**	Non-linguistic element	*· § % …*
79	**EM**	Emoticon	*:-)*

Table 1: Tagset.

	full tagset	UD UPOS	Google UPOS
Ann1 vs. Ann2	90.63	92.83	94.07
Ann1 vs. Gold	91.89	93.67	94.72
Ann2 vs. Gold	98.64	99.12	99.33

Table 2: Agreement between annotators and gold standard for the full tagset and for versions mapped to the coarser tagsets. Values are accuracy percentages.

4. Mapping to Universal Part-of-Speech Tags

For some applications and use-cases, it is useful to have a more coarse-grained set of part-of-speech tags that makes a broad distinction between word-classes but abstracts away from most of the additional properties encoded in the tagset described above. Rather than invent a new coarse-grained set of labels, we adopt the popular Universal Part-of-Speech Tags that have become a quasi standard. There are two flavors of Universal Part-of-Speech Tags:

- Google UPOS, the original Google Universal Part-of-Speech Tagset by Petrov et al. (2012) that consists of 12 tags[2] and

[2] https://github.com/slavpetrov/universal-pos-tags

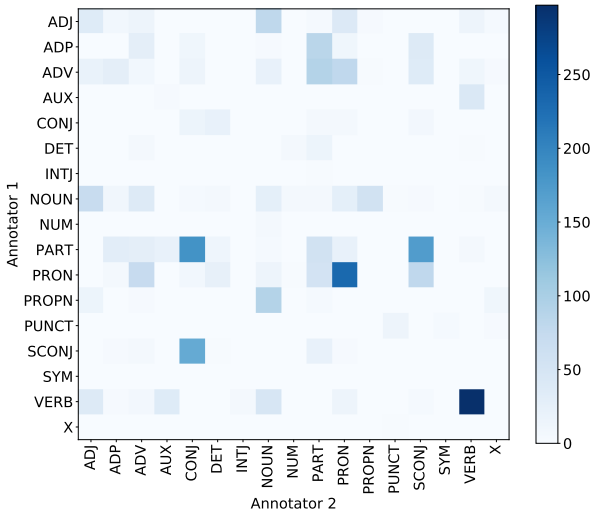

Figure 1: Tagging differences between the annotators. For clarity of presentation, we map the tags to UD UPOS tags, i. e. entries on the diagonal indicate differences between tags that get mapped to the same UD UPOS tag.

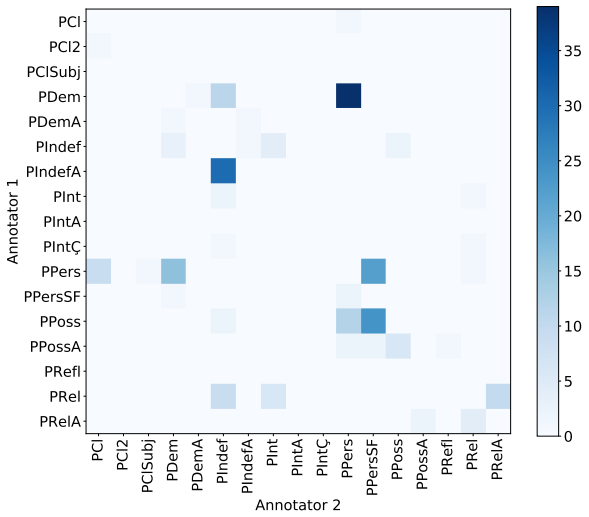

Figure 2: Tagging differences between the annotators (pronominal tags only).

- UD UPOS, the refined version used in the treebanks of the Universal Dependencies project (Nivre et al., 2016) that consists of 17 tags.[3]

In Table 3, we provide a mapping from our tagset to both varieties of the Universal Part-of-Speech Tagset.

The mapping is straightforward. All of our tags can be seen as refinements of the Universal Part-of-Speech Tagset – a deliberate design choice.

5. Evaluation

For our evaluation, we perform tagging experiments using the following part-of-speech taggers:

[3] http://universaldependencies.org/

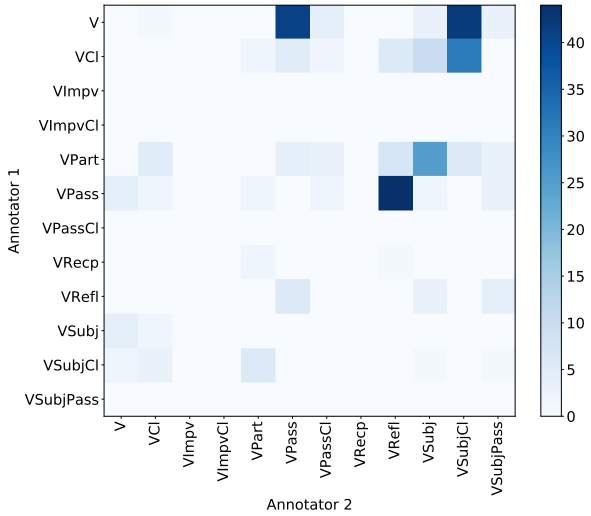

Figure 3: Tagging differences between the annotators (verbal tags only).

- The HMM-based HunPos tagger (Halácsy et al., 2007),[4] an open source reimplementation of Brants' TnT tagger (Brants, 2000).
- The maximum entropy tagger from the Apache OpenNLP project[5].
- TreeTagger (Schmid, 1994; Schmid, 1995),[6] a tagger based on decision trees.
- SoMeWeTa (Proisl, 2018),[7] a perceptron-based tagger that can make use of external resources. We provide it with Brown clusters (Brown et al., 1992) extracted from 82 million tokens of Albanian text.
- The Stanford POS Tagger (Toutanova et al., 2003),[8] a maximum entropy tagger that uses cyclic dependency networks.

We evaluate tagging accuracy in five scenarios using ten-fold cross-validation on our gold standard corpus. The five scenarios are:

1. Training and testing using our full tagset,
2. training and testing using UD UPOS,
3. training and testing using Google UPOS,
4. training using our full tagset and mapping the output to UD UPOS for testing and
5. training using our full tagset and mapping the output to Google UPOS for testing.

The results are shown in Table 4. At first glance, they seem to be rather modest. Using the full tagset, the best tagger achieves 91.00% accuracy – a good six points less than the state of the art for languages like English, French or German. However, we have to take into consideration that our corpus is orders of magnitude smaller than for example the Wall Street Journal part of the Penn Treebank and that it is at the

[4] https://code.google.com/archive/p/hunpos/

[5] https://opennlp.apache.org/

[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[7] https://github.com/tsproisl/SoMeWeTa

[8] https://nlp.stanford.edu/software/tagger.html

Google	UD	tags
ADJ	ADJ	Adj, AdjA, AdjP, AdjPA, AdjN, NumO
ADP	ADP	Prep
ADV	ADV	Adv, AdvPt, AdvMP
CONJ	CONJ SCONJ	ConjC, ConjC1, ConjC2 ConjS, ConjS1, ConjS2
DET	DET	Art
NOUN	NOUN PROPN	N, NA, NHg NE
NUM	NUM	NumC
PRON	PRON	PPers, PPersSF, PDem, PDemA, PPoss, PPossA, PInt, PIntA, PIntÇ, PRel, PRelA, PIndef, PIndefA, PRefl, PCl, PCl2, PClSubj
PRT	PART	Pt, PtComp, PtCond, PtFut, PtGer, PtInf, PtJus, PtMod, PtNeg, PtNegD, PtNegS, PtPass, PtPassCl, PtPriv, PtProg, PtProh, PtQA, PtQM, PtSubj
VERB	AUX VERB	VAux, VMod V, VPart, VCl, VImpv, VImpvCl, VPass, VPassCl, VSubj, VSubjCl, VSubjPass, VRecp, VRefl
X	INTJ SYM X	Intj EM, NLE Abbr, FW
.	PUNCT	Punct, Punct2

Table 3: Mapping table from our tagset to Universal Part-of-Speech Tags (both the Google and the Universal Dependencies variant).

same time much more heterogeneous. In addition, Albanian has a much richer morphological system than English and our tagset is more than 50% larger than that of the Penn Treebank.

Training the taggers using the full tagset and mapping the predicted tags to one of the UPOS tagsets eliminates within-class errors produced by the taggers and leads to much better results with up to 94.68% accuracy. For almost all taggers, this setting works better than training directly on one of the coarser tagsets, i. e. the taggers benefit from a more fine-grained internal representation. The exception to this rule is SoMeWeTa, the only tagger provided with additional external knowledge in the form of Brown clusters, which achieves better results of up to 95.10% when trained directly on the coarser tagsets.

The main sources of errors for SoMeWeTa on the full tagset are visualized in Figure 4. The largest group of errors is confusion between different verb tags. Next are confusions between different noun tags (mostly between N and NHg) and, to a lesser extent between verb and noun tags and between different tags for pronouns. The most frequent misclassifications for the 20 tags with the most errors are shown in Table 5. The single most difficult distinction for the tagger is that between N and NHg, i. e. between nouns and heterogeneous nouns. This is not surprising given that

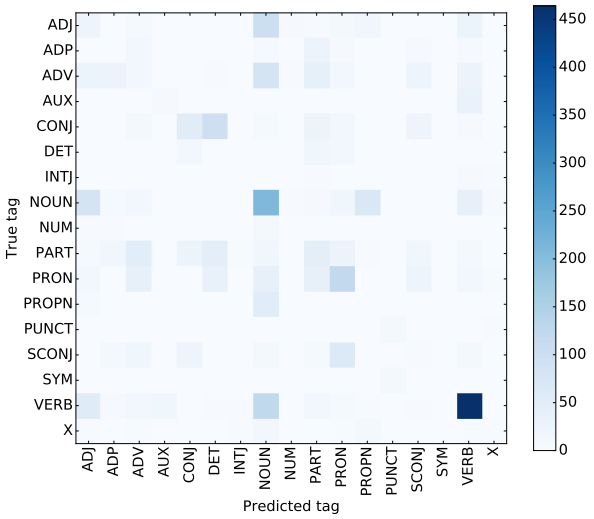

Figure 4: Errors made by SoMeWeTa on the full tagset. For clarity of presentation, we map the tags to UD UPOS tags, i. e. errors on the diagonal indicate misclassifications between tags that get mapped to the same UD UPOS tag.

this is a lexical distinction that cannot be deduced from context. To be able to correctly distinguish between the two, the tagger must have seen the word during training or must be provided with a corresponding lexicon. Other frequent misclassifications include noun (N) vs. proper noun (NE), noun (N) vs. adjective (ADJ) and coordinating conjunction (ConjC) vs. article (Art).

6. Conclusion and Future Work

Up to now, there has been no manually tagged corpus of substantial size for Albanian. The gold standard presented in this paper is by far the largest manually tagged corpus for Albanian and consists of 2,020 sentences (31,584 tokens) annotated with a medium-sized part-of-speech tagset designed with a focus on the syntagmatic aspects of the language, especially multi-word units involving articles or particles. While the corpus is still too small to achieve state-of-the-art tagging accuracies comparable with those for better-resourced languages, the evaluation experiments show very promising results similar to what could be expected from an English corpus of similar size. For the coarser tagsets, we achieve accuracies of up to 95.10%. We are also optimistic that the results could be further improved by providing the taggers with an additional lexicon. Such a lexicon could be derived from one of the existing morphological analyzers.

7. Bibliographical References

Arkhangelskiy, T., Belyaev, O., and Vydrin, A. (2012). The creation of large-scale annotated corpora of minority languages using UniParser and the EANC platform. In *Proceedings of COLING 2012: Posters*, pages 83–92. The COLING 2012 Organizing Committee.

Brants, T. (2000). TnT – A statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, pages 224–231.

Brown, P. F., Pietra, V. J. D., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of

tagger	full tagset	UD UPOS	Google UPOS	mapped to UD UPOS	mapped to Google UPOS
HunPos	88.73 ±0.79	92.44 ±1.18	93.63 ±1.18	92.72 ±0.86	94.04 ±0.78
SoMeWeTa	**91.00** ±0.83	**94.32** ±1.40	**95.10** ±1.38	**93.99** ±0.97	**94.68** ±0.97
TreeTagger	88.14 ±1.12	91.15 ±1.69	92.15 ±1.87	92.24 ±1.20	93.43 ±1.17
OpenNLP	87.50 ±1.23	89.87 ±1.39	91.24 ±1.35	90.64 ±1.35	91.98 ±1.20
Stanford Tagger	85.96 ±1.26	89.72 ±1.28	91.37 ±1.24	89.92 ±1.59	91.36 ±1.49

Table 4: Evaluation results (mean accuracy percentages ±2 standard deviations).

tag	freq	err	most frequent confusions
N	5470	233	NE (63), NHg (45), Adj (35), V (25)
Adv	965	198	N (67), ConjS (21), Prep (20), Pt (20)
V	1562	178	N (62), VRefl (32), VCl (22), VSubj (18)
ConjC	1234	172	Art (98), ConjS (19), Pt (16), PCl (10)
NHg	493	155	N (141), NE (3), VRefl (3), Adj (2)
VRefl	401	150	V (56), VPass (35), VSubj (19), N (11)
ConjS	656	133	PRel (56), ConjC (20), Adv (18), N (10)
Pt	292	131	Adv (25), ConjC (20), ConjS (15)
Adj	904	112	N (69), NE (10), V (8), AdjA (8)
VCl	486	82	V (14), N (13), AdjA (13), VRefl (11)
VPass	142	82	VRefl (57), V (14), VPart (3), PtPass (2)
VSubj	376	70	VPart (15), AdjA (14), VSubjCl (9)
PIndef	308	69	Adv (20), N (19), PIndefA (8), PInt (4)
PCl	694	68	Art (29), PtPass (14), PtSubj (9)
NE	712	63	N (54), Adj (3), AdjA (3), PCl (1)
Prep	2292	63	Adv (14), PtInf (10), PtComp (10), V (5)
NA	124	57	AdjA (28), N (16), PIndefA (5), VCl (2)
AdjA	1118	54	N (17), NA (7), VCl (6), NE (5)
VPart	526	52	N (20), AdjA (8), VSubjCl (8), VCl (3)
Art	2996	46	PtSubj (16), PCl (11), ConjC (11)

Table 5: The 20 tags that SoMeWeTa most frequently misclassified.

natural language. *Computational Linguistics*, 18(4):467–479.

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

Demiraj, S., 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.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Kabashi, B. and Proisl, T. (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 2016)*, pages 4305–4310, Paris, France. European Language Resources Association.

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

Kabashi, B. (2017). AlCo – një korpus tekstesh i gjuhës shqipe me njëqind milionë fjalë. In *Seminari XXXVI Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare*. Universiteti i Prishtinës, Prishtinë.

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

Leech, G. and Wilson, A. (1999). Standards for tagsets. In Hans van Halteren, editor, *Syntactic Wordclass Tagging*, pages 55–80. Kluwer Academic Publishers, Dordrecht.

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

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Paris, France. European Language Resources Association.

Petrov, S., Das, D., and McDonald, R. T. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, May 23-25, 2012*, pages 2089–2096.

Piton, O. and Lagji, K. (2008). Morphological study of Albanian words, and processing with NooJ. In *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 *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007)*, pages 407–413. Springer, Berlin, Heidelberg.

Proisl, T. (2018). SoMeWeTa: A part-of-speech tagger for German social media and web texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki. European Language Resources Association.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop*, pages 47–50, Dublin.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Marti A. Hearst et al., editors,

Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003).

Trommer, J. and Kallulli, D. (2004). A morphological analyzer for standard Albanian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1271–1274. European Language Resources Association.

8. Language Resource References

Péter Halácsy and András Kornai and Csaba Oravecz. (2007). *HunPos*. `https://code.google.com/archive/p/hunpos/`.

Thomas Proisl. (2017). *SoMeWeTa*. `https://github.com/tsproisl/SoMeWeTa`.

Apache OpenNLP project team. (2004). *Apache OpenNLP*. `https://opennlp.apache.org/`.

Helmut Schmid. (1994). *TreeTagger*. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`.

Kristina Toutanova and Dan Klein and Christopher Manning and William Morgan and Anna Rafferty and Michel Galley and John Bauer. (2004). *Stanford Log-linear Part-Of-Speech Tagger*. `https://nlp.stanford.edu/software/tagger.html`.