PREPROCESSING TOOLS FOR THE ALBANIAN LANGUAGE: A STATE-OF-THE-ART SURVEY AND AN ANNOTATION SCHEMA PROPOSAL

Nelda KOTE
Faculty of Information Technology, Polytechnic University of Tirana, Albania
Marenglen BIBA
Faculty of Information Technology, New York University of Tirana, Albania

ABSTRACT
Language preprocessing tools as parser, morphological analyzer, and lemmatizer have become very popular Natural Language Processing (NLP) approaches. Firstly, in this paper, we have presented a literature review of recent researches on preprocessing tools for the Albanian language. There is a lack of works done to implement these tools for the Albanian language. To our knowledge, there is not an official published preprocessing tool for the Albanian language. Secondly, in this paper, we have proposed a part-of-speech annotation schema for the Albanian language. We mapped the annotation schema proposed by Salavaçi and Biba (2012) into a new annotation schema containing the part-of-speech tags and morphological characteristics tags. Improving the proposed annotation schema and developing a morphological tagger for the Albanian language will foster.
Keywords: POS tagger; stemming; lemmatize; Albanian language

1. INTRODUCTION

Nowadays, there are tools as part-of-speech (POS) tagger, stemmer, and lemmatizer for almost all languages. Efforts have been made to develop these tools for the Albanian Language, but to our knowledge, no one of these tools is available online and public for use. Despite the lack of NLP tools, there is a lack of high-quality public annotated corpora for NLP purposes in the Albanian Language.

This paper aims to evaluate the state-of-the-art situation of POS taggers, lemmatizers, and stemmers in the Albanian Language and to propose a POS annotated corpora for the Albanian Language. We have proposed an annotated schema based on the Universal Dependencies schema. The

proposed annotated corpus using this schema is created by mapping an existing annotated corpora into the new schema.

The Albanian language is the official language in the Republic of Albania and the Republic of Kosovo, the second official language in the Republic of North Macedonia, and the official regional language in Ulcinj, Montenegro. It is an isolated branch of the Indo-European language family. The Albanian language has a complex morphology and syntax system. There are two dialects, Tosk used by the standard Albanian, and Gheg. The Albanian alphabet is Latin and has 36 letters, 29 consonants and 7 vowels. The standard Albanian language morphology has ten different word categories: noun, pronoun, verb, adverb, numeral, particle, conjunction, preposition, and interjection. The morphological categories of the noun are the case (nominative, genitive, dative, accusative, and ablative), definiteness (defined and non-defined), gender (female, masculine and neutral), and number (singular and plural). Some masculine nouns in plural number form change the gender to female and are known as endogen nouns. The adjectives can be: linked with an article or not. The adjective adapts in case, number, and gender with the word they modify. So, the adjective has the morphological categories of case, number, and gender. There are seven types of pronouns, possessive, interrogative, demonstrative, subject, relative, indefinite, and reflexive. The verbs have the most complex inflection system. The morphological categories of the verb are the person (1, 2, 3), number (singular and plural), voice (active and passive), mood (indicative, admirative subjunctive, imperative and optative), tense (present, past, and future). The adverbs have the morphological category of gradation. There are five types of adverbs of quantity, manner, time, cause, and location. The numbers, particle, conjunction, preposition, and interjection are non-changeable words and do not have morphological categories. The complex inflectional paradigms of the Albanian language make the development of NLP tools difficult. The forthcoming section discusses the efforts made to create annotated corpora and develop NLP tools for the Albanian Language.

The structure of the paper is as follows: Section 2 present a literature review of the NLP tools for English; Section 3 presents the literature review of the part-of-speech and stemmer tools of Albanian; Section 4 presents a proposal for an annotation schema for the Albanian language, and in Section 4 we conclude our work.

2. Literature review: Natural Language Processing (NLP) Tools

Natural Language Processing (NLP) tools as a part-of-speech tagger, stemmers, and lemmatizers are linguistic procedures performed in computational linguistic throw a wide range of algorithms. A part-of-speech tagger aims to assign to each word in a sentence a morphosyntactic class

based on lexical and contextual information. Stemmer and lemmatizer aim to transform the word respectively into its root (stem) and lemma. These tools can be rule-based or statistical. The rule-based algorithms are developed based on a set of language-dependent morphological rules. Language dependency is one of the most important disadvantages of these algorithms. This disadvantage has led the researchers to focus on developing statistical techniques. Statistical techniques eliminate the language dependency problem, and the developed tools can be used for different languages. These algorithms use supervised or unsupervised statistical techniques. The disadvantage of supervised techniques is the need to have annotated corpora. The creation of an annotated corpus is time-consuming and requires specialized persons to do it. The performance of these tools depends significantly on the languages they are used for. More complex languages require powerful techniques to have good performance in POS tagging, stemming, and lemmatization.

In 1992, Brill (1992) developed the first and the most widely used rule-based part-of-speech tagger for English. This tagger assigns to each word a tag based on the Penn Treebank tagset. The Pen Treebank tagset has 36 tags for morphological annotation of English. The method used by this tagger required a reduced store space, and the identification of the errors is easy, but the tagging process is slow.

The supervised technique is used to develop the TnT tagger (Brants, 2000) and the Stanford Log-linear POS Tagger (Toutanova *et al*., 2000). The TnT tagger (Brants, 2000) is an implementation of the Viterbi algorithm for the second-order Hidden Markov Model (HMM) model. As scholastic approaches, the HMM model and Viterbi algorithm assign the most probable tag to a word when used to POS tagging. This tagger is implemented in English and German using three different corpora with an accuracy of 96.7%. The Stanford Log-linear POS Tagger (Toutanova et al., 2000) is a POS tagger based on a maximum entropy algorithm. Overthrow the years it is improved and implemented in different languages. This model enhances the performance of tagging by enriching the information sources used to tag. The authors have implemented the POS tagger using the Penn Treebank with an accuracy of 96.86% and 86.91% for unseen words.

The first developed stemmer algorithm is rule-based. In 1968, Lovins (1968) developed the first context-sensitive, longest-match rule-based stemming algorithm for English. The algorithm has two phases. In the first phase, the longest matched ending of the word from a predefined list of 294 endings is removed. Then in the second phase, the stem is treated to solve some linguistic exceptions like double consonants or irregular plurals utilizing one of the 35 transformation rules. This stemmer is very fast, but too many rules needed to be implemented. As the algorithm has a limited number of predefined endings, there can be a high number of errors (Lovins, 1968).

Dawson (1974) improved and extended Lovin's stemmer by using a list of 1200 endings and using the partial matching technique to match stems that are equal within certain limits. Dawson's stemmer algorithm uses a large set of endings and requires more time and storage to be executed.

Porter's stemmer is one of the most used rule-based stemming algorithms in Information Retrieval. The first version of this algorithm was for English, and later due to its good results, it was used for languages like German, French, Russian, etc. This algorithm has five steps, uses around sixty suffixes, has two rules responsible for recording and one context-sensitive rule to decide if a suffix should be removed or not. In 2001, was created the Snowball framework (Poter, 2001) that includes an improved version of the algorithm for English, German, French, Russian, etc.

In 2013, Mayfield and McNamee (2003) proposed the first language-independent stemming algorithm. This N-gram stemming algorithm offers the possibility to develop models for different languages. The experimental results demonstrate the effectiveness of the selection of a single n-gram as a stem of a word.

Last years, researchers are using more and more neural networks or deep learning in the development of parsers, part-of-speech taggers, stemmers, or lemmatizes. The models based on these technologies offer the possibility that one proposed solution to be easily used in different languages. Different tools like these have been implemented and evaluated in the framework of a lot of shared tasks.

3. Literature review: annotated corpora for the Albanian language

In this section, we present the existing annotated corpora and morphological tagging tools for the Albanian Language. The grammar of the Albanian language is complex, and the process of the annotation is a challenge. The complicated inflectional paradigm of the Albanian language makes the development of stemming and lemmatization tools too difficult. There are some attempts to develop rule-based stemming tools that we have discussed in detail below.

Trommer and Kallulli (2004) proposed a simple rule-based morphological tagger for the standard Albanian language taking into consideration the main grammatical rules of the Albanian language. The tagger uses 340 morphological rules that indicate the relation between input lexicon and output derived forms. The tagset is a pair of sets attribute-value conform to EAGLE guidelines standard adapted to the Albanian language grammar. The tagset contains 17 labels: n,v, a, prsp, part, reflp, posp, demp, indp, intp, relp, pa, prep, adv, ptl, seq, conj. The tagger is evaluated in two small corpora, each of 500 tokens (word) so we cannot define if this is a performing tool.

The tagger proposed by Salavaçi and Biba (2012) may be considered as the first statistical part-of-speech tagger for Albanian language using the OpenNLP tool. The authors have annotated a corpus of 10.000 words using three tagsets, respectively with 100 tags, 150 tags, and 220 tags. We have discussed in detail this tagset in section 3. This annotated corpus is used to train two models, a Maxent model, and a Perceptron model. The average accuracy of the two models is nearly 60%. We will use this annotated corpus to create the annotated corpora using the Universal Dependencies (UD) schema.

The Saint-Petersburg linguists' team (Morozova and Rusakov, 2014) created The Albanian National Corpus contains 16.6 million tokens. This is the biggest annotated corpus for the Albanian language. The corpus contains text from short stories, novels, fiction and non-fiction memoirs, journals, religious official, and scientific texts. The tagset used have 62 tags which include the standard tags as a verb, adverb, preverb, adjective, adjective numeral, numeral, pronominal clitic, conjunction, preposition, prepositional article, particle, interjection, pronoun, and other tags about the case, gender, number, definiteness, mood, person, tense, verbal representation, animacy, transitivity, voice, pronominal clitics, and article position. Each word has the correspondent lemma, the English translation, and the morphological features of the lexeme and the word form. The authors report that based on four Albanian dictionaries they have manually created the files that contain a grammatical wordlist with all the information of the lexeme and the inflectional paradigm. The corpus is morphologically tagged automatically by using the morphological analyzer UniParser, which assigns separately to each word of the corpus a tag considering the information in these files (Arkhangelskiy et al., 2012). The process of tagging does not take into consideration the syntactic context of the word form when determining its grammatical features. The authors do not report the performance of the Albanian model in terms of accuracy. The tagged model is not available online to be downloaded; the annotated corpus is available online, but cannot be downloaded to be used for NLP purposes.

Kadriu (2013) developed a morphological tagger using the NLTK toolkit. To assign part-of-speech tags to new text this morphologic tagger uses a corpus annotated corpus of 32000 words and a set of regular expressions rules. In the first phase, a tokenizer is used to tokenize the sentences and the words, then only nouns and verbs are assigned the lemma by removing or replacing the suffixes. In the second phase, the word is annotated with the part-of-speech tags using the NLTK. First, the word is tagged using the tagger based on the dictionary and then is used the regular expression tagger determines the correct tag to a word by using a series of regular expressions and affixes. The tagset contains 22 tags. In case there is a word not

determined in the dictionary or cannot be generated by the regular expression pattern, it is tagged with the tag None.

The implemented model has an accuracy of up to 90% but the corpus used to train, and to test is small, and the tagger generated a considerable number of untagged words as the words non-found in the dictionary, the irregular verbs, nouns, and adjectives.

Kirov et al. (2018) have created a small morphological annotated corpus for the Albanian language. The corpus contains 589 lemmas, and 33483-word forms in total using the UniMorph schema. To each word is assigned the corresponding lemma and the morphological tags. The trained models and the annotated corpus are free downloaded.

Kabashi and Proisl (2018) have created a morphologically annotated corpus for the Albanian language, but it is not available online. The corpus contains 2020 sentences, 31584 tokens. Two native Albanian speakers have manually annotated it. They presented three tagsets, one full tagset specified by the authors, and the corresponding mapping tagset to Google UPOS and Universal Depends UPOS. The full tagset has in total 79 tags classified in 16 main tags categories: 4 tags for nouns, 14 tags for verbs, 5 tags for adjectives, 3 tags for adverbs, 14 tags for pronouns, 1 tag for preposition, 6 tags for conjunctions, 2 tags for numbers, 19 tags for particles, 1 tag for interjections, 1 tag for articles, 3 tags for pronominals, 2 tags for abbreviations, 2 tags punctuations, 1 tag for emoticon and 1 tag the non-linguistic element. Even they have used a large tagset they have not included tags for the number, gender, definiteness, and case of the name. The authors have specified a large number of tags to distinct nouns and adverbs preceded by an article. The annotated corpus is used to develop five models using SoMeWeTa tagger, HMM-based HunPos tagger, TreeTagger, Stanford POS Tagger, and OpenNLP tagger. The accuracy of the models is 85.96% to 95.10%, and SoMeWeTa is the best performant model.

Karanikolas (2009), a non-native Albanian speaker, developed the first rule-based stemming algorithm for the Albanian language.

This algorithm works on the principle of the longest-match suffix removal. The author, based on five Albanian grammatical books, has defined a list of 470 stopwords. The suffix remover is done in three steps, firstly is removed the longest suffix that matches the ending of the word, then the generated word is used to remove another matched longest suffix, and in the final step is kept only the first of a sequence of multiple ending consonants from the word generated by the second step. This algorithm is manually evaluated in a small corpus.

Sadiku and Biba (2012) presented the JStem algorithm, the rule-based stemming algorithm implemented in Java Programming for the Albanian language. For the Albanian language, this can be considered the first

stemming algorithm developed by a native speaker. The algorithm uses 134 rules that remove the suffixes and prefixes of a word to generate its root and a stopwords list. The authors have not specified rules for plural formation, feminine, masculine, and neutral gender formation. The JStem algorithm is extended by Biba and Gjati (2014) to stem the composite words in the Albanian Language. By analyzing the morphology structure of composite words in the Albanian language, the authors have added a set of rules to find the stems of the composite words. The implemented rules take into consideration the composite words formed by two words linked by a separating dash, with prefixes, with numerals, by productive nouns associated with verbs or adjectives, with international prefixes, and formed randomly by associating two different words.

4. A proposal for an annotation schema and an annotated corpus for the Albanian language

This section presents a grammar annotation schema (part of speech and morphological features) for the Albanian language using the Universal Dependencies (UD) schema. We have considered the tagset and the annotated corpus developed by Salavaçi and Biba (2012) and propose a new annotated corpus based on the rules of the UD schema.

The annotated corpus of Salavaçi and Biba (2012) does not have the lemmas of the words, so we have manually assigned to each word the corresponding lemma. To define the correct form of the lemma, we have used the Albanian National Dictionary.

The existing tagging set is mapped to the new tagging set that follows the UD schema. One of the problems we fronted in this process is that tagging in the UD schema is word-level, while Salavaçi and Biba (2012) in the annotation consider the difference between an adjective preceded or not by an article and the verbs in a composite tense are annotated as an entity and not every single part of it separately.

The corpus annotation using the new schema was done automatically by a program developed by us considering the mapping rules we defined for the old and the new annotation schema.

The tagset presented by Salavaçi and Biba (2012) has 224 tags in total. There are 39 tags for nouns, a tag for particular nouns, and 38 for nouns indicating gender, number, case, and definitiveness. They have not considered if the noun is preceded by an article or not as defined in the Albanian language grammar. We use the tag NOUN to tag a noun and PROPN to tag a person's name. To nouns, we have assigned the morphological features of the case, definite, gender, and number. We have not assigned morphological tags to a person's names.

Salavaçi and Biba's (2012) tagset specifies 20 different tags for the article, indicating the gender, number, and case. In our schema, we have used the DET tag and no morphological feature.

Table 1. Noun annotation examples.

	Salavaçi and Biba's (2012)		Our Annotation	
	Annotation	Tag Meaning		
derën	derën_NNFNSK	Noun - Female - Singular - Definitive - Accusative	NOUN	Case=Acc Definite=Def Gender=Fem Number=Sing
Veliu	Veliu_NNP	Particular noun	PROPN	

One of the most complex word categories in the Albanian language is the verb. Salavaçi and Biba (2012) have used 75 tags for the verb. There is a tag for the compound verb tense to tag the verb kam (have) and jam (is) used in the Albanian language as an auxiliary verb to form the compound tenses. For this category, they have not specified any grammatical characteristics. They have different tags for modal verbs, negative verbs, infinitive, gerundive, and participle forms of the verb. The other 69 tags are used to tag the verbs based on the different morphological features, like mood, tense, person, number. In this tagging schema, a verb in a compound tense is tagged as the auxiliary verb with tag Vb+ and the verb with the corresponding tag that represents the mood, tense, person, and the number of this compound verb. We have adopted this schema to the UD schema using only the main tag VERB and the AUX tag for the mund (can) verb. As morphological features, we use case, number, person, and tense. For the verb in compound tense, we have used the tag VERB even for the auxiliary verb and the main verb and the same grammatical features tags for the two verbs. For the infinitive form of the verb, we have used the morphological feature VerbForm=Inf. For the gerund form of the verb, we have used the morphological feature VerbForm= Ger. For the participle form of the verb, we have used the morphological feature VerbForm= Part. Maybe our verb tagging schema will need to be revised to a better proposal. Table 2 shows some examples.

Table 2 Verb annotation examples.

	Salavaçi and Biba's (2012)		Our Annotation		
	Annotation	Tag Meaning			
kishte bërë	kishte_VB+ bërë _VBDMIIIN	VB+ -> Verb - Compound verb VBDMIIIN - > Verb - Indicative - Past Perfect - Person III - Singular	kishte	VERB	Mood=Ind Number=Sign Person=3 Tense=Past
			bërë	VERB	Mood=Ind Number=Sign Person=3 Tense=Past
pat kthyer	pat_VB+ kthyer_VBDKTIIIS	VB+ -> Verb - Compound verb VBDKTIIIS -> Verb - Indicative - Pluperfect - Person III – Plural	pat	VERB	Mood=Ind Number=Plur Person=3 Tense=Pqp
			kthyer	VERB	Mood=Ind Number=Plur Person=3 Tense=Pqp
do të ketë	do_RP të_RP ketë_VBDAIIIN	PR -> Particle VBDAIIIN -> Verb - Indicative - Future - Person III – Singular	do	PART	
			të	PART	
			ketë	VERB	Mood=Ind Number=Sing Person=3 Tense=Fut

Salavaçi and Biba (2012) have used 68 tags for the pronouns specifying the type and the morphological features of gender, number, and the case for the demonstrative, personal, interrogative, and indefinite pronouns. They have a different tag for subjunctive pronoun, a reflexive pronoun, short form of the personal pronoun, personal pronoun contraction short form, possessive pronoun, agglutinated pronouns, short for of personal pronoun, and indefinite form of the personal pronoun. For these pronouns' categories, they have not defined morphological features. We have tagged all the pronouns with the PRON tag. The PronType morphological feature is used to define the type of pronouns. To the demonstrative pronouns, we have assigned the morphological features of the case, gender, number, and prontype with the value Dem. The morphological features used to personal pronouns are the case, gender, number, person, and prontype with the value Prs. To the short form of the personal pronoun, personal pronoun contraction short form, and

agglutinated pronoun, we have defined the morphological features pron type with value Prs. To the pronouns tagged by Salavaçi and Biba (2012) with the tag possessive pronoun, we have assigned the morphological features Poss=Yes and PronType=Prs. We tagged the interrogative pronouns with the morphologic features case, gender, number, and prontype with the value Int. And the indefinite pronouns are tagged with the morphologic features case, gender, number, and prontype with the value Ind.

Table 3 Pronouns annotation examples.

	Salavaçi and Biba's (2012)		Our Annotation	
	Annotation	Tag Meaning		
ata	ata_PRVIIIMSK	Pronoun Personal - Person III - Masculine - Plural - Accu.	PRON	Case=Acc Gender=Masc Number=Plur Person=3 PronType=Prs
kjo	kjo_PRiPFNE	Indefinite Pronoun - Female - Singular - Nominative	PRON	Case=Nom Gender=Fem Number=Sing PronType=Dem
tim	tim_PRPRN	Possessive pronoun	PRON	Poss=Yes PronType=Prs
pse	pse_PRP	Interrogative Pronoun	PRON	PronType=Int

Salavaçi and Biba (2012) have defined 6 different tags for the adverbials, one for each category in the Albanian language. We have used the ADV tag to tag an adverbial and the morphological feature AdvType to define the adverbial type.

Table 4 Adverbial annotation examples.

	Salavaçi and Biba's (2012)		Our Annotation	
	Annotation	Tag Meaning		
si	si_ABM	Adverbial - Clauses of manner	ADV	AdvType=Man
përse	përse_ABQ	Adverbial - Purpose clauses	ADV	AdvType=Cau

In the Albanian language, the conjunction is a non- inflected word category, and Salavaçi and Biba (2012) used only one tag. We have used the CCONJ and the SCONJ tag and no morphological feature.

To tag numbers, Salavaçi and Biba (2012) have used three tags, one for cardinal numbers and two for ordinal numbers specifying their gender female or masculine. We use the NUM tag for numbers and the morphological feature NumType with value Ord for ordinary numbers and with value Card for cardinal numbers.

In the Albanian language, the preposition is a non-inflected word that fits in the case with the noun or adjective that follows them. Salavaçi and Biba (2012) have used four tags for prepositions considering their case. The preposition in our schema is tagged as ADP, and we have defined the morphological feature of the case.

In the Albanian language, the adjective is inflected for the case, number, and gender. In addition, it can be either articulated or unarticulated. Moreover, the adjectives have the degrees of comparison. The tagset of Salavaçi and Biba (2012) defined 5 tags for this word-class indicating whether it is articulated or not, or gradability (affirmative, comparative, and superlative degrees). The adjective is tagged using the ADJ tag, and no morphological features are assigned. Based on Albanian grammar, we need to add the morphological features of the case, number, and gender in the future.

We have used the PUNCT tag for punctuation and INTJ tag for the interjection and onomatopoeic words.

Table 5 Annotation examples

	Salavaçi and Biba's (2012)		**Our Annotation**		
	Annotation	Tag Meaning			
dhe	dhe_CC	Conjunction	CCONJ		-
kur	kur_CC	Conjunction	SCONJ		-
një	një_CDRF	Ordinal numeral - Female	NUM		NumType=Card
dy	dy_CDRM	Ordinal numeral - Masculine	NUM		NumType=Card
1935	1935_CDT	Cardinal numeral	NUM		NumType=Ord
nga	nga_INE	Preposition - Nominative	ADP		Case=Nom
e thellë	e_CC thellë_JJNP	CC -> Conjunction JJNP-> Adjective - With Article - Base	e	CCOJN	-
			thellë	ADJ	Degree=Pos
shumë prekës	shumë_RP prekës_JJPS	RP -> JJPS-> Adjective - Without Article - Superlative	shumë	PART	-
			prekës	ADJ	-
.	._.	punctuation	PUNCT		-

The Figure 1 illustrates an example of an annotated sentence by Salavaçi and Biba (2012) and Table 6 shows the annotated sentence using our proposed tag schema.

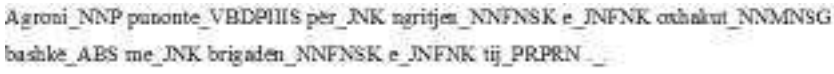

Fig. 1: Salavaçi and Biba sentence annotation.

Table 6 Annotated sentence using our proposed tag schema

Word	Lemma	POS tag	Morphological features
Agroni	Agron	PROPN	
punonte	punoj	VERB	Mood=Ind Number=Plur Person=3 Tense=Past
për	për	ADP	Case=Acc
ngritjen	ngritje	NOUN	Case=Acc Definite=Def Gender=Fem Number=Sing
e	e	DET	
oxhakut	oxhak	NOUN	Case=Gen Definite=Def Gender=Masc Number=Sing
bashkë	bashkë	ADV	AdvType=Man
me	me	ADP	Case=Acc
brigadën	brigadë	NOUN	Case=Acc Definite=Def Gender=Fem Number=Sing
e	e	DET	
tij	tij	PRON	Poss=Yes PronType=Prs
.	.	PUNCT	

5. CONCLUSION

In the present a survey of the existing morphological annotation tools and annotated corpus for the Albanian language is presented. There is no official tool to be used for the Albanian Language. We have proposed a new annotation schema for the Albanian language by mapping the annotation schema of Salavaçi and Biba (2012) to a new one based on UD Schema. The annotated corpus in (Salavaçi and Biba 2012) is mapped to the new annotation schema by an automatic program. The annotation is the first step and the most important to develop modern morphological annotation tools. In the Albanian language, the annotation process is challenging due to its grammar complexity. The annotation schema in the present paper proposed can be the first step to create a stable annotation schema for the Albanian language. The annotation corpus also can be the first step towards an annotation corpus based on a multilingual annotation schema. In the future, we plan to enlarge the annotated corpus and use it to develop and implement a morphological tagger.

REFERENCE

Arkhangelskiy T, Belyaev O, Vydrin A. 2012. The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform. Proceedings of COLING 2012. 83-91.

Biba M, Gjati E. 2014. Boosting text classification through stemming of composite words. Recent Advances in Intelligent Informatics, Advances in Intelligent Systems and Computing. 235: 185-194.

Brill E. 1992. A simple rule-based part of speech tagger. ANLC '92 Proceedings of the third conference on Applied natural language processing. 152-155.

Brants Th. 2000. TnT: a statistical part-of-speech tagger. ANLC '00 Proceedings of the sixth conference on Applied natural language processing. 224-231.

Dawson J. 1974. Suffix removal and word conflation. *ALLC Bulletin*. **2(3):** 33-46.

Kabashi B, Proisl Th. 2018. Albanian part-of-speech tagging: Gold standard and evaluation. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018. 2593–2599.

Kadriu A. 2013. NLTK Tagger for Albanian using Iterative Approach. Proceedings of the 35th International Conference on Information Technology Interfaces, 2013. 283-288.

Karanikolas N. 2009. Bootstrapping the Albanian Information Retrieval. Fourth Balkan Conference in Informatics, 2009. 231-235.

Kirov Ch, Cotterell R, Sylak-Glassman J, Walther G, Vylomova E, Xia P, Faruqui M, Mielke S, McCarthy A, Kubler S, Yarowsky D, Eisner J, Hulden M. 2018. UniMorph 2.0: Universal Morphology. LREC.

Lovins J. 1968. Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics. 11.

Mayfield J, McNamee P. 2003. Single n-gram stemming. SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 415-416.

Morozova M, Rusakov A. 2014. Albanian national corpus: Composition, text processing and corpus-oriented grammar development, language and culture of the Albanian. The 5th Deutsch-Albanischen Cultural Science Conference.

Porter MF. 2001. Snowball: a language for stemming algorithms.

Sadiku J, Biba M. 2012. Automatic Stemming of Albanian Through a Rule-based Approach. *Journal of International, Research Publications: Language, Individuals and Society*, **6:** 173-190.

Salavaçi E, Biba M. 2012. Enhancing Part-of-Speech Tagging in Albanian with Large Tagsets.

Toutanova K, Manning DCh. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 63-70.

Trommer J, Kallulli D. 2004. A Morphological Tagger for Standard Albanian. Proceedings of LREC.