Computational Modeling of Morphologically Rich Languages – The Case of Nouns in Albanian Language

Anila Çepani, University of Tirana, Albania
Adelina Çerpja, Academy of Sciences of Albania, Albania

The Barcelona Conference on Education 2024
Official Conference Proceedings

Abstract
The Albanian language is a synthetic-analytical one. It's rich system of inflection poses significant challenges in developing computational models for morphological analysis. Our primary question is how to construct a computational morphological model specifically for the nominal system in Albanian, focusing on nouns which have various grammatical categories and forms, including number (singular and plural), gender (masculine, feminine, neutral), case (nominative, genitive, dative, accusative and ablative) and definite and indefinite forms. The initial step involves analysing the morphological structure of the nominal system of the Albanian language, identifying its grammatical categories, forms, and means of construction, such as endings, inflectional suffixes, stem alternations, suppletion, and combinations. To address this challenge, an exact methodology was employed, involving the development of formulas based on different noun stems to encompass all possible forms for each grammatical variation. These formulas were crucial for generating different forms, aiming to minimize manual intervention and streamline the automatic completion of nominal forms. The subsequent step is evaluating the developed models by comparing their results to manually constructed forms, improving the accuracy and efficiency, and validating the effectiveness of the models by testing them in real applications, such as Albanian spell checking. These models are indispensable for building applications for spelling and grammar in Albanian, as well as other NLP applications, enhancing natural language processing tools for the Albanian language.

Keywords: Albanian Language, Computational Modeling, Morphology, Nouns, Inflection, Grammatical Features

1. Introduction

In today's technology-driven environments, Natural Language Processing (NLP) has become an indispensable element, serving as a cornerstone in numerous fields and applications. However, the effectiveness of NLP systems heavily significantly relies on their capability to accurately process languages with complex morphological structures. Automatic parsing of morphologically rich languages presents difficulties due to the interaction of diverse morphemes within words. Unlike languages with simpler morphological structures, such as English, morphologically rich languages develop various morphemes to express different grammatical features. This complexity presents obstacles to parsing algorithms, as they must accurately segment words into constituent morphemes and determine their grammatical roles. In recent years, Albanian Natural Language Processing has made significant strides, spurred by increasing interest in computational linguistics and the demand for tailored language technologies. Notable advancements include corpus development: National Corpus of the Albanian Language emerging as one of the most extensive resources (Arkhangelskij et al., 2012); Besim Kabashi's creation of the AlCo corpus (Kabashi, 2017); Nebi and Ali Caka's CAKA corpus (Caka & Caka, 2012); and the sqGLOBE corpus (BFSU 2008-2012) developed by Chinese researchers. Morphological and syntactic analysis have been central areas of focus, with scholars such as Trommer and Kallulli (2004), Caka and Neziri (2011), and Kabashi and Proisl (2018) developing taggers and analyzers to address challenges like morphological ambiguity. Efforts in Named Entity Recognition (NER) have been spearheaded by researchers including Skënduli and Biba (Sadiku & Biba, 2012) and Kote et al. (2019), leveraging machine learning methodologies and Conditional Random Fields (CRF) to enhance performance in this domain Additionally, sentiment analysis studies by Biba and Mane (2014) and Ajdari et al. (2017), among others, have explored emotion detection and hate speech identification in Albanian text. These endeavours underscore progress in Albanian NLP, although challenges persist, notably the scarcity of annotated data and the necessity for more extensive corpora.

Throughout the development of the "Albanian Language in the Digital Era" project, led by the Center for Educational and Promotion in Pristina (https://gjuhashqipe.com) and supported by the Ministry of Education, Science, and Innovation in Kosovo, our focus has been on constructing a computational morphological model encompassing all parts of speech and their various forms in Albanian. The focus of this paper is the model designed for the nominal system of Albanian language, which targets nouns that present a diverse range of grammatical categories and forms, covering singular and plural numbers, masculine, feminine, and neuter genders, as well as cases such as nominative, genitive, dative, accusative, and ablative, alongside definite and indefinite forms.

Albanian language belongs to the Indo-European family and is a rich morphology language, especially the verbal and noun system. Today, the Albanian language is spoken by more than seven million people in the Republic of Albania, the Republic of Kosovo, the western and northwestern parts of the Republic of North Macedonia, some municipalities in the southeastern part of the Republic of Montenegro, the municipalities of Preševo, Bujanoc, and Medvedja (Republic of Serbia), some areas of Greece (Chameria, etc.), and the old Albanian settlements of Italy, Greece, Ukraine, Bulgaria, etc., as well as in Albanian communities in various countries worldwide.

This paper provides an overview of techniques for analysing morphologically rich languages, focusing on Albanian nouns, characterized by various grammatical features, including

gender, number, case, and definiteness, each expressed through a multitude of form patterns. Developing computer models for noun morphology is essential for various natural language processing tasks, including parsing, lemmatization, spell checking, text generation, machine translation, information retrieval etc. We describe the process of creating digital models to generate and analyse Albanian noun forms, describing the complexities and challenges encountered in this process.

The first step is to analyse the morphological structure of the nominal system of the Albanian language, identifying its grammatical categories, forms, and means by which these forms are constructed. An important point in this process is data collection, which includes gathering a wide range of nouns and all their possible grammatical features and variations. All these forms are used to construct the development of computational models of nouns in Albanian, based on the collected data and devised formulas about the morphological transformations of nouns. The next step is the evaluation of the developed models, comparing their results to manually constructed forms, and improving the accuracy and efficiency, and finally, the validation of the effectiveness of the models by testing them in real applications, such as Albanian spell checker. The first version of the Albanian spell-checker took place as part of the "Gjuha shqipe dhe kompjuteri" project (1998–2004), organized by the Qendra për Edukim dhe Përparim (QEP) in collaboration with linguists and computer scientists from Albania and Kosovo (Çepani & Çerpja, 2017: 89-106).

2. General Knowledge About the Nominal System of the Albanian Language

The grammatical categories of the nouns are gender, number, case, and the definiteness/indefiniteness (Agalliu et al., 2002).

Grammatical Category of the Gender of the Noun

The gender is an important grammatical category of the noun and every noun in the Albanian language can be in one of the three genders: masculine, feminine and neuter. Most nouns are masculine or feminine; only a small number of nouns are neuter.

Gender is an inherent feature of nouns and is syntactically independent in Albanian. The ending sounds of the stems and the case endings of the definite and indefinite forms are a morphological way, which serve to distinguish the gender of nouns together with the syntactic and lexical way.

- The masculine definite forms end in *-i* or *-u*: *emër – emr**i*** (noun)*; çaj – çaj**i*** (tea)*; kek – kek**u*** (cake)*; zog – zog**u*** (bird).
- The feminine definite forms end in *-a* or *-ja*: *kuzhinë – kuzhin**a*** (kitchen)*; vezë – vez**a*** (egg)*; shije – shij**a*** (taste)*; gjyshe – gjysh**ja*** (grandmother)*; domate – domat**ja*** (tomato).
- The neuter definite forms end in *-t(ë)* or *-it: të ngrënë**t*** (eating)*; të kuq**të*** (redness); *të gatuar**it*** (cooking)*; të shijuar**it*** (tasting).

Grammatical Category of Number to Noun

In the Albanian language, the noun has two numbers: *singular* and *plural*. Most nouns have special forms for both numbers: *mur – mur**e*** (wall)*, çantë – çant**a*** (bag)*,* etc. But there are some singularia tantum nouns in Albanian, which are used only in singular forms, e.g. *sheqer* (sugar)*, benzinë* (gasoline); or proper nouns, such as *Albania, Saranda,* etc., while some other

nouns are used only in the plural number, such as the pluralia tantum: *pantallona* (pants), *syze* (glasses) etc.

Singular and plural of the nouns in Albanian are distinguished by:

a. The plural suffixes or stem alternations of the noun itself: *mal – mal**e*** (mountain/s), *libër – libr**a*** (book/s), *mik – mi**q*** (friend/s), *breg – br**igje*** (cost/s).
b. Noun modifiers that agree the noun in number: *laps i zi – lapsa të zinj* (black pencil – black pencils), *ai laps – ata lapsa* (that pencil – those pencils); *shtëpi e re – shtëpi të reja* (new house – new houses), *kjo shtëpi – këto shtëpi* (*this house – these houses*).

The Albanian nouns have a special stem for the plural, to which the endings of indefinite or definite plural cases are appended. The stem of singular can be the same with the stem of plural for a group of nouns: *nxënës – nxënës* (pupil/s), *fletë – fletë* (sheet/s), while for the rest of the nouns these two stems are different and the plural stem is formed from the singular one by other means.

The plural stem is formed in one of these four ways:

1. **Without any means,** ie. the stem of the plural is the same as the stem of the singular: *mësues – mësues* (teacher/s), *lot – lot* (tear/s), *mollë – mollë* (apple/s), *nxënës – nxënës* (pupil/s), *përkthyes – përkthyes* (translator/s), *kafshë – kafshë* (animal/s), *shtëpi – shtëpi* (house/s), etc.;
2. **By plural suffixes:** *mal – mal**e*** (mountain/s), *fshat – fshat**ra*** (village/s), *kumbull – kumbull**a*** (plum/s), *kopsht – kopsht**e*** (garden/s), *drejtor – drejtor**ë*** (director/s)**,** *hero – hero**nj*** (hero/es)**,** *lumë* **-** *lum**enj*** **(**river/s), etc.;
3. **By vowels** or **consonants changes:** *dorë – d**ua**r* (hand/s), *plak – pl**e**q* (old man/men), *grua – gr**a*** (woman/women)**,** *peshk – pesh**q*** (fish/es)**,** *bir – bi**j*** (son/s), etc.;
4. **By plural suffixes** and **vowels** or **consonants changes at the same time:** *shteg – sht**igje*** (path/es), *rrezik – rrezi**qe*** (risk/s)**,** *bllok – bllo**qe*** (block/s)**,** *lëng – lën**gje*** (liquid/s), etc.

Grammatical Category of Case and Definiteness / Indefiniteness Form

Nouns in the Albanian language can be used in different cases forms according to their syntactic function in the sentence. There are five cases in Albanian: nominative, genitive, dative, accusative, and ablative. Definiteness is expressed using inflectional suffixes.

The nominative case has two forms: an indefinite and a definite form. The defined form is created by adding the suitable case marker (m., f., n.) to the indefinite form.

Some of the forms are the same in different cases, i.e:

*Kjo **shkollë** është e re.* (nominative case)
This **school** is new.

*Po lyejnë **shkollën**.* (accusative case)
They are painting the **school**.

The genitive forms differ from the forms of the dative and ablative only by the preceding article:

*Sot është ditëlindja **e shokut** të klasës*. (genitive case)
Today is my classmate's **birthday.**

*Librin ia dhurova **shokut** të klasës*. (dative case)
I gave the book to my **classmate.**

*Këtë dhuratë e kam prej **shokut** të klasës*. (ablative case)
I got this gift from my **classmate.**

The formation of the five cases forms in Albanian nouns is complicated and follows different rules. These forms are used for different syntactic purposes and often require additional morphological changes to the noun stem.

The totality of all the changes that the noun undergoes when it is used in different cases is called *declination.* These changes differ according to the case, and each of them comes out in the definite and indefinite form. The nouns in the Albanian language are grouped into four declinations according to the ending of the definite nominative case of the singular form:

Declination I includes all masculine nouns that end in *-i* in singular, definite nominative case, e.g. *djal-**i*** (son)**,** *lis-**i*** (oak)**,** *burr-**i*** (man)**,** *fto-**i*** (quince)**,** *vëlla-**i*** (brother), etc.

Declination II includes masculine nouns that end in *-u* in singular, definite nominative case, e.g. *mik-**u*** (friend)*, zog-**u*** (bird)*, dhe-**u*** (earth)**,** *ah-**u*** (beech), etc.

Declination III includes feminine nouns that end in *-a, -ja* in singular, definite nominative case*,* e.g. *vajz-**a*** girl*, tryez-**a*** (table)*, fush-**a*** (field)*, motr-**a*** (sister)*, lul-**ja*** (flower)*, del-**ja*** (sheep), etc.

Declination IV includes neuter nouns that end in *-it, - t(ë),* in singular, definite nominative case, e.g. *të folur**it*** (speaking), *të ftohtë-**t*** (the cold), *të ri-**të*** (youth), etc.

3. Modeling the Nominal System of the Albanian Language

The morphological structure of the nominal system in the Albanian language, involving various grammatical categories, forms, and means of construction, such as endings, inflectional suffixes, stem alternations, suppletion, and combinations of them, necessitates the development of computational models to generate and analyse noun forms. An exact methodology is used to construct such models, involving the development of formulas based on different noun stems to encompass all possible forms for each grammatical variation. These formulas are crucial for generating different forms, aiming to minimize manual intervention and to simplify the automatic completion of nominal forms.

Table 1: Models of the Albanian Nominal System

Gender	Declination	Plural	Suffixes and types of sound change	Examples
masculine	I	as singular	as singular	*nxënës* (pupil)
			as singular (-Ë)	*çallmëbardhë* (white turban)
			Abbreviation (i)	*KESH*
		with suffixes	e	*gabim* (error), *rreth* (circle)
			e -Ë	*cikël* (cycle)
			a -Ë	*libër*
			-a	*bel* (spade), *plep* (poplar)
			-ë	*anëtar* (member), *luftëtar* (fighter)
			-nj	*hu* (stake), *kalli* (cob)
			-inj	*drapër* (hook)
			-inj	*shkëmb* (rock)
			-j + N extension	*kalama* (kiddy)
			-j + R extension	*kufi* (limit)
			-enj	*lumë* (river)
			-ër	*mbret* (king)
			-ra	*bar I* (grass)
			-na	*bar II* (drug)
		with sound changes	palatalization ll>j	*akull* (ice), *avull* (steam)
			palatalization r>j	*bir* (son), *lepur* (rabbit)
			palatalization ll> j and first vowel change	*mashkull* (male)
			palatalization ll>j and second vowel change	*bakall* (grocer)
		with sound changes and suffixes	vowel change ua>o + -NJ	*ftua* (quince)
			palatalization ll>j, vowel change + -E	*huall*(honeycomb), *truall* (land)
		other sound changes	a>e	*cjap* (goat)
			e>a	*Thes* (sack), *rreth* (circle)
			i>e	*vit-vjet* (year)
			vowel change ë>u + -Ë	*dhëndër* (bridegroom)
			palatalization g>gj + -E	*gardh* (fence)
		with one big difference	with a partially same or different composition MANUAL	*djalë – djem* (boy), *kalë - kuaj* (horse)
	II	with suffixes	acronim (u)	*KEK* (KEK)
			-ë	*lek* (lek), *ibrik* (kettle)
			-nj	*borxhli* (debtor)
			-ra	*gjak* (blood)
		with sound changes	palatalization (k>q)	*mik* (friend)
			palatalization g>gj)	*zog* (bird)
		with sound changes and suffixes	palatalization k>q, vowel change a>e + -E	*bllok* (block)
			palatalization g>gj) +-E	*lëng* (liquid)
			palatalization g>gj, vowel change e>i + -E	*breg* (shore)
			palatalization k>q, vowel change a>e + -E	*lak* (loop)
		with suppletive form	with a partially same or different composition manual	*ka - qe* (cow)
feminine	III	as singular	as singular -A	*këngë* (song)
			as singular -JA	*nxënëse* (schoolgirl)
			as singular -JE	*mbledhje* (meeting)

			as singular -IE	*borxhlie* (indebted)
			as singular -E	*vegane* (vegan)
			as singular accented (i)	*shtëpi* (house)
			as singular accented (a, e)	*sevda* (love)
			as singular -ja (o)	*depo* (storehouse)
			acronym (definite)	*UEFA* (UEFA)
			acronym (-së)	*ATSH* (*ATSH*)
			acronym (-s)	*NATO* (*NATO*)
		with suffixes	-a	*fushë* (field)*, verë II* (summer)
			-a (..ël)	*vegël* (tool)
			-a (..ull)	*kumbull* (plum)
			-a article -ë	*e drejtë* (right)
			-a nyjë	*e çelur* (blown)
			-ra	*verë I* (wine)
			composite f. with fem. noun	*dhëmbësharrë* (tooth saw)
			composite m. with fem. noun	*dhëmbësharrë* (tooth saw)
		with one big difference	with a partially same or different composition	*grua* - *gra* (woman)
neuter	**IV**	**without amount**	nominal participle -IT	*të folurit* (speaking)
			nominal participle -ËT	*të qarët* (crying)
			nominal neutral adjective	*të nxehtët* (hot)

3.1. Data Collection

The first step in modelling the nominal system of the Albanian language starts by collecting a comprehensive dataset of nouns and their corresponding forms, including variations in gender, number, case, and definiteness / indefiniteness. This dataset serves as the foundation for developing computational models and ensuring their accuracy and coverage across different linguistic contexts.

3.2. Formula Development

A rule-based approach was employed to address the complex nature of noun inflection while developing digital morphological models for Albanian nouns. This methodology included a systematic analysis of the patterns of Albanian noun paradigm, followed by the formulation of rules to generate all possible forms of a noun. As we mentioned above, nouns in the Albanian language have several grammatical categories, each characterized by a number of forms. Every noun has different forms according to the gender, number, case and the definite or indefinite form. Each of these grammatical categories affects the inflectional patterns associated with the noun.

Formulas are developed from the collected data to systematically generate noun forms. These formulas are valid for all morphological transformations of nouns, considering the variations in gender, number, case, and definiteness / indefiniteness. By implementing these formulas within computational models, it becomes possible to automate the generation of most noun forms without manual intervention, thereby enhancing efficiency and scalability.

According to the data table above and comparing the nominal and verbal system in Albanian, it's clear that, although the noun has fewer grammatical categories than the verb, it has many patterns even within the same grammatical category.

The challenge of designing this algorithm for nouns is the generation of all noun forms in the Albanian language, which begins with the selection of the declination, and one of the respective patterns of the nouns for each declination. There are three stems about each noun (see Table 2: *nxënës* - pupil), on which the generation of all forms of this noun and the corresponding explanations are performed. The formulas for generating these nouns are used for all other nouns that are included in the same group as "nxënës", which have the same forms as for the singular and for the plural, for example *mësues* (teacher), *shitës* (seller), etc.

Table 2: Basic Stems for the Noun *NXËNËS* (pupil)

The masculine noun NXËNËS (*pupil)*		
The formula	**Grammatical features**	**Stems**
E1	nominative, singular, indefinite	nxënës
E2=E1	nominative, singular, definite	nxënës
E3=E1	nominative, plural, indefinite	nxënës

This algorithm is very important, considering that the formulas of the noun *nxënës* serve as an exact pattern to automatically generate all nominal forms for 10.477 nouns in Albanian.

Several steps must be followed for the operation of this algorithm, which are related to the different stems of the noun. First, the user must choose the type of noun considering the declination and the changes it undergoes in the plural number. This helps the algorithm to automatically complete the three stems:

1. E1 - the singular stem of the nominative case, singular, indefinite form, is automatically filled in, it's the representative form of the nouns in Albanian.
2. E2 - the stem of the nominative case, singular, definite form changes in some nouns and does not change in others. It does not change in the case of noun *nxënës* (pupil) E2 = E1.
3. E3 - the stem of the nominative case, plural, indefinite form, differs in a considerable number of nouns. This stem is also used for all definite forms in the plural, and it is E3 = E1 in the case of noun *nxënës* (pupil).

Table 3: The Full Forms of Masculine Noun NXËNËS (pupil)

Indefinite, singular			**Definite, singular**	
nominative	një	*nxënës*	nominative	*nxënësi*
genitive (i...)	një	*nxënësi*	genitive (i...)	*nxënësit*
dative	një	*nxënësi*	dative	*nxënësit*
accusative	një	*nxënës*	accusative	*nxënësin*
ablative (prej...)	një	*nxënësi*	ablative (prej...)	*nxënësit*
Indefinite, plural			**Definitive, plural**	
nominative	ca	*nxënës*	nominative	*nxënësit*
genitive (i...)	ca	*nxënësve*	genitive (i...)	*nxënësve*
dative	ca	*nxënësve*	dative	*nxënësve*
accusative	ca	*nxënës*	accusative	*nxënësit*
ablative (prej...)	ca	*nxënësish*	ablative (prej...)	*nxënësve*
	ca	*nxënësve*		

Table 4: The Forms With Formula of Masculine Noun NXËNËS (pupil)

Indefinite, singular			**Definite, singular**	
nominative	një	*E1*	nominative	*E1+i*
genitive (i...)	një	*E1+i*	genitive (i...)	*E1+it*
dative	një	*E1+i*	dative	*E1+it*
accusative	një	*E1*	accusative	*E1+in*
ablative (prej...)	një	*E1+i*	ablative (prej...)	*E1+it*
Indefinite, plural			**Definite, plural**	
nominative	ca	*E1*	nominative	*E1+it*
genitive (i...)	ca	*E1+ve*	genitive (i...)	*E1+ve*
dative	ca	*E1+ve*	dative	*E1+ve*
accusative	ca	*E1*	accusative	*E1+it*
ablative (prej...)	ca	*E1+ish*	ablative (prej...)	*E1+ve*
	ca	*E1+ve*		

Starting from the way different nouns construct the form of the nominative case, the singular, as well as the form of the plural, including the difference in gender, we have compiled 65 representative models of formulas for the automatic generation of different nouns forms in the Albanian language, which are illustrated below with the nouns representing the four declination in Albanian: TRUALL (land), ZOG (bird), KËNGË (song), and TË QARËT (crying).

Table 5: The Masculine Noun TRUALL (land) – 1st Declination

The formula	**Grammatical features**	**Stems**
E1	nominative, singular, indefinite	truall
E2=E1	nominative, singular, definite	truall
E3=E1+e (ua>o, ll >j)	nominative, plural, indefinite	troje

Table 6: The Masculine Noun ZOG (bird) – 2nd Declination

The formula	**Grammatical features**	**Stems**
E1	nominative, singular, indefinite	zog
E2=E1	nominative, singular, definite	zog
E3=E1 (g>gj)	nominative, plural, indefinite	zogj

Table 7: The Feminine Noun KËNGË (song) – 3rd Declination

The formula	**Grammatical features**	**Stems**
E1	nominative, singular, indefinite	këngë
E2=E1-ë	nominative, singular, DEFINITE	këng
E3=E1	nominative, plural, indefinite	këngë

Table 8: The Neuter Noun *TË QARËT* (crying) – 4th Declination

The formula	**Grammatical features**	**Stems**
E1	nominative, singular, indefinite	qarë
E2=E1	nominative, singular, definite	qarë
E3=E1-ë	nominative, plural, indefinite	qar

3.3. Model Evaluation

These developed computational models have undergone to a rigorous evaluation to assess their accuracy and effectiveness in generating noun forms. This evaluation involved comparing the output of the models against manually constructed forms, identifying any inconsistency or errors, and refining the models accordingly. By continuous improving the models based on evaluation results, their overall performance and reliability could be enhanced.

3.4. Application Testing

Once validated through evaluation process, the computational models were tested in real-world applications to demonstrate their utility and effectiveness. This testing phase involved integrating the models into NLP tools for Albanian language processing, such as spell checkers, grammatical analysers. These applications served as e very good and important way of enhancing the quality during the evaluating process about the performance of these models (Çerpja & Çepani, 2022).

4. Challenges in Computer Modeling

The challenge in the computer modelling of nouns in the Albanian language is related to the correct completion of their different forms, considering the interaction of the grammatical categories of nouns. Determining the appropriate inflectional forms for each noun means choosing them according to the gender and number.

In building algorithms for noun morphology, it is essential considering the different inflectional patterns associated with masculine, feminine, and neuter nouns, as well as the irregularities and exceptions within them. The presence of irregular nouns and exceptions complicates the modelling process. Irregular nouns may deviate from standard inflection patterns, requiring special attention within computer models.

Likewise, exceptions to standard inflection rules introduce inaccuracies during automatic generation, so they require attention in modelling formulas to ensure accurate generation of noun forms, but the number of ‘irregular’ nouns in Albanian language is very low compared to all the nouns.

5. Conclusions

Modelling the nominal system of the Albanian language presents significant challenges due to its rich morphological structure and diverse grammatical categories. However, by applying rigorous methodologies for data collection, formula development, model evaluation, and application testing, there are constructed computational models capable of generating and analysing noun forms accurately and efficiently. These models are essential for various NLP tasks in Albanian language processing, facilitating the development of advanced linguistic tools and applications.

The digital morphological structure has served as the basis for the construction of Albanian spelling and grammar, as an integral component of the "Albanian Language in the Digital Era" project. This integration has highlighted the accuracy of the generation of nouns forms.

The effectiveness of these models can be further enhanced through continued research and refinement, particularly in handling irregular cases and dialectal variations, contributing to the advancement of NLP technologies for morphologically rich languages like Albanian.

References

Agalliu, F. et al. (2002). *Gramatika e gjuhës shqipe*, I, Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, Tiranë.

Arkhangelskij, T., Daniel, M., Morozova, M., & Rusakov, A. (2012). Korpusi i gjuhës shqipe: drejtimet kryesore të punës // *Shqipja dhe gjuhët e Ballkanit. Albanian and Balkan Languages*. Konferencë e mbajtur më 10–11 dhjetor 2011 në Prishtinë / red. Rexhep Ismajli. Prishtinë: ASHAK. 635–642.

Biba, M., & Mane, M. (2014). Sentiment analysis through machine learning: An experimental evaluation for Albanian. In *Recent Advances in Intelligent Informatics.* 195–203.

Bolshakov, I. A., & Gelbukh, A. (2004). *Computational Linguistics Models, Resources, Applications*, 186 pp; ISBN 970-36-0147-2, www.gelbukh.com/clbook

Caka, A., & Neziri, V. (2011). Algoritmi i modelit kompjuterik të etiketuesit (tagerit) të gjuhës shqipes (Computer model Algorithm of the Albanian language tagger). *Alb-Shkenca – Konferenca e Seksionit të Shkencave Inxhinierike dhe të Teknologjisë së Informacionit*.

Caka, N., Caka, A. (2012). Korpusi i gjuhës shqipe – rezultatet e para, problemet dhe detyrat. *Shqipja dhe gjuhët e Ballkanit – Albanian and Balkan Languages.* Prishtinë, 643–656.

Çeliku, M., Domi, M., Floqi, S., Mansaku, S., Përnaska, R., Prifti, S., & Totoni, M. (2002). *Gramatika e gjuhës shqipe. Vëllimi II Sintaksa.* Akademia e Shkencave e Republikës së Shqipërisë. Tiranë.

Çepani, A., & Çerpja, A. (2017). *Hyrje në gjuhësinë kompjuterike* (tekst universitar). Fakulteti i Historisë dhe i Filologjisë, "Albas", Tiranë, ISBN 978-9928-02-833-4.

Çerpja, A., & Çepani, A. (2022). Shqipja në erën digjitale −arritje dhe perspektiva. In *Akte të Kuvendit Ndërkombëtar të Studimeve Albanologjike, Vëllimi II*. 407-437, Akademia e Shkencave e Shqipërisë, Akademia e Shkencave dhe e Arteve e Kosovës, ISBN 978-9928-339-73-7. Tirana.

Collaku, I., & Adal, E. (2015). Morphological parsing of Albanian language: a different approach to Albanian verbs. In *International Conference on Computer Science and Communication Engineering.* 87–91, doi:10.33107/ubt-ic.2015.94

Çollaku, I., & Adali, E. (2015). Morphological parsing of Albanian language: a different approach to Albanian verbs. *UBT International Conference*. 94. https://knowledgecenter.ubt-uni.net/conference/2015/all-events/94

Gjuha letrare shqipe për të gjithë. Elemente të normës letrare kombëtare. Komisioni hartues: Prof. Androkli Kostallari (kryetar), Emil Lafe, Menella Totoni, Nikoleta Cikuli. Shtëpia Botuese e Librit Shkollor. Tiranë, 1976, 294 f.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science.* vol. 349, no. 6245, 261–266, doi: 10.1126/science.aaa8685

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, see www.cs.colorado.edu/~martin/slp.html

Kabashi, B. (2017). ALCO – një korpus tekstesh i gjuhës shqipe me njëqind milionë fjalë. *Seminari ndërkombëtar për Gjuhën, letërsinë dhe kulturën shqiptare*. Prishtinë, 36, 123–132.

Kabashi, B., & Proisl, T. (2018). Albanian part-of-speech tagging: Gold standard and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* 2593–2599.

Kastrati, M., & Biba, M. (2022). Natural language processing for Albanian: a state-of-the-art surve. *International Journal of Electrical and Computer Engineering (IJECE).* Vol. 12, No. 6, 6432–6439.

Kostallari A., Lafe E., Totoni, M., & Cikuli, N. (1976). *Gjuha letrare shqipe për të gjithë. Elemente të normës letrare kombëtare*. ShBLSh. Tiranë, 294 f.

Kote, N., Biba, M., Kanerva, J., Rönnqvist, S., & Ginter, F. (2019). Morphological Tagging and Lemmatization of Albanian: A Manually Annotated Corpus and Neural Models. *CoRR abs/1912.00991*.

Rregullat e drejtshkrimit të shqipes (Projekt) (1967). Komisioni hartues: Androkli Kostallari, Eqrem Çabej, Mahir Domi, Emil Lafe (asistent). Universiteti Shtetëror i Tiranës, Instituti i Gjuhësisë dhe i Letërsisë. Tiranë, 214 f.

Sadiku, J., & Biba, M. (2012). Automatic stemming of Albanian through a rule-based approach. *Journal of International Research Publications: Language, Individuals and Society.* vol. 6.

Shishani, L., & Çerpja, A. (2005). Gjuha shqipe dhe programi për drejtshkrim AS 2.0. *Gjuha jonë.* n. 1–4, f. 126–134.

Trommer, J., & Kallulli, D. (2004). A morphological tagger for standard Albanian. In *Proceedings of LREC.* 1–8.

Zenuni, X., Ajdari, J., Ismaili, F., & Raufi, B. (2017). Automatic hate speech detection in online contents using latent semantic analysis. Pressacademia, 5(1), 368–371. https://doi.org/10.17261/pressacademia.2017.612

Contact email: anila.cepani@unitir.edu.al