INNOVATION IN LANGUAGE LEARNING International Conference

COMPUTATIONAL MODELING OF MORPHOLOGY IN ALBANIAN LANGUAGE: THE CASE OF VERBS

Adelina Çerpja [1], Anila Çepani [2]

[1] Institute of Linguistic and Literature, Academy of Albanian Studies, Tirana, Albania
[2] Faculty of History and Philology, University of Tirana, Tirana, Albania

Abstract

The Albanian language is synthetic-analytical and, as a language with developed inflection, it has a rich system of grammatical forms. To prepare applications for spelling and grammar in a language, as well as several NLP applications, the development of computer models of morphological forms is particularly important. This process in the case of the Albanian presents many difficulties and challenges. In this paper, we describe the process of creating a computational morphological model of the verbal system in the Albanian language. The verb in Albanian has the grammatical categories of person, number, tense, mood, and diathesis. The grammatical meanings of these categories are expressed with an exceptionally considerable number of grammatical forms, which are constructed with different means, which serve to express grammatical meanings: personal endings, alternations of the stem of the verb, inflectional suffixes, suppletion, and/or combinations between these. To create digital morphological models of verbs in the Albanian language and to assign morphological labels and lemmas, it was necessary to prepare different formulas based on different stems of the verbs, which serve to generate all verb forms for each mood, tense, person, number, etc. These are a small group of inductive and representative models that, despite the structural complications and diverse means of verb forms, result in the most accurate and automatic completion of the forms for each verb in the Albanian language.

Keywords: *Albanian language, software, digitalization, morphology, verb*

1. Introduction

The Albanian language forms a separate branch in the family of Indo-European languages. It is a synthetic-analytical language, with a predominance of synthetic features and a tendency towards analyticity. Albanian is considered a language with developed inflection and consequently has a rich system of grammatical forms, especially for nouns and verbs.

During the development of successful models for their analysis a special importance is given to morphological analysis in terms of preparing applications for proofreading and editing in a certain language and in several natural language applications.

Morphological analysis is essential and forms a core subsystem for other NLP applications. The morphological level is concerned with several tasks that focus on the internal structure of words and how they are realized in language. These tasks are otherwise called the normalization of the text, which means converting it into a more suitable and standard form.[1]

This paper describes the process of creating computational morphological models of verbal system in Albanian language, considering that the many Albanian morphological forms pose special challenges to computational natural language

[1] D. Jurafsky; J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, p. 10.

processing systems.[2] This work is part of the project "Albanian language in the digital era", which is carried out by the Center for Educational and Promotion (https://gjuhashqipe.com), Prishtina, supported by the Ministry of Education, Science, and Innovation in Kosova.

To create computer morphological models of verbs in the Albanian language, to determine morphological labels and lemmas, it was necessary to prepare different formulas based on verb stems, which serve to generate all verb forms for every mood, tense, person, number etc. These are inductive models that, regardless of structural complications and the variety of means and verb forms, result in the most accurate and automatic completion of the forms for each verb in the Albanian language.

2. General knowledge of the verbal system of the Albanian language

The verb in the Albanian language is characterized by several special grammatical categories,[3] as the category of person, number, mood, time and diathesis, which are expressed by special endings, inflectional particle, and auxiliary verbs (*kam* and *jam*).

The inflectional forms of the verbs in Albanian are synthetic and analytical.

Synthetic forms of the verb are built: by endings (*la* -j, *la* -n); by phonetic changes (*dal: del, njoh: njeh*); by personal endings and phonetic changes (*dal: dol-a, thye-j: the-va*); by inflectional suffixes, sometimes with personal endings (*la-fsh-a, la-rë*); by suppletive forms with personal endings or inflectional suffixes (*jam: qe-shë: qe-në*).

Analytical forms of the verb are built: by the auxiliary verbs *kam* and *jam* used before the participle of the main verb (kam *larë,* jam *larë*); by inflectional particles (duke *larë,* u *lava*).

The Albanian verbs are grouped into three conjugations, divided into classes and subclasses, and the irregular verbs.

3. Representative models of the verbs in the Albanian language

CONJUGATION I (Verbs ending in the consonant *j)*	**Class I**	***Subclass 1***	*punoj, rrëfej, shkruaj, lyej*
		Subclass 2	*blej*
	Class II	***Subclass 1***	*arrij, mbaj, ruaj*
		Subclass 2	*bëj*
CONJUGATION II (Verbs ending in a consonant)	**Class I**	***Subclass 1***	*hap, mas*
	Class II	***Subclass 1***	*dredh, nxjerr, pjek, djeg, dal, marr*
		Subclass 2	*përkas*
		Subclass 3	*shkas*
CONJUGATION III (Verbs ending in a vowel)	**Class I**		*vë*
	Class II		*di*
	Class III		*pi*
	Class IV		*shtie*
IRREGULAR	**Class I**		*jam*
	Class II		*them*

[2] This is an important part of the project "Albanian language in the digital era" of Center for Education and Promotion – QEP (https://gjuhashqipe.com/fillimi), funded by the Ministry of Education, Science, Technology and Innovation in Kosovo.

[3] For this general knowledge, we are based on the chapter of the verb in *Grammar of the Albanian language*, I, Academy of Sciences of Albania, Institute of Linguistics and Literature, Tirana, 2002.

4. Wordform generation formulas for each pattern and statistic

As we noted above, the verb in the Albanian language has different grammatical categories and a variety of forms for each category. Each of the moods of the verb has a certain number of tenses which depending on the formal aspect, are simple tenses **(***synthetic*) and compound tenses (*analytical).*

We should note that some verb forms are the same in different persons, tenses, and moods. The number of forms about one verb is 480, but the number of non-repetitive forms varies from 429 to 432.

The challenge of designing this algorithm is the generation of all verb forms in the Albanian language, which starts with the selection of the verb class and continues with the automatic generation of verb forms based on the models with which this algorithm is equipped. An important process here is the determination of the different stems of a verb, since, as Baerman points out[4], stems and the changes they undergo behave similarly to affixes and are therefore evaluated with the same parameters as affixes.

We are giving below the different stems of a verb (*mësoj-learn*), on which the generation of all forms of this verb is performed, and the corresponding explanations.

Table 0-1: **Basic stems for the verb MËSOJ (learn).**

The verb MËSOJ (*learn*)		
Formula	**Grammatical features**	**Stems**
F1=headword -j	present I, singular	*mëso*
F2=F1	present II, singular	*mëso*
F3=F1	present II, plural	*mëso*
F4=F1	simple past I, singular	*mëso*
F5=F1 (o>ua)	simple past I, plural	*mësua*
F6=F5+r	participle	*mësuar*

Several steps must be followed for the operation of this algorithm, which are related to the different stems of the verb:

1. The user chooses the type of verb considering the changes it undergoes in the present tense, simple perfect and participle.
2. F1 - the first-person singular stem of the present indicative, is automatically filled in.
3. F2 - the stem of the second person singular of the present indicative, changes in some verbs and doesn't change in others, and in this case F2 = F1.
4. F3 - the stem of the second person plural of the present indicative, differs in a considerable number of verbs. This stem is also used for all passive forms built with the respective endings.
5. F4 - the stem of the singular simple past tense is one of the stems that most often undergoes changes, either in vowels or in consonants, and for many verbs it serves as the stem of other forms, such as the simple forms of the optative and admirative mood.
6. F5 - the stem of the plural simple past tense, often is the same as F4, but there are verbs in which it undergoes changes, so it is left as a separate stem.
7. F6 - the participle, is an important stem, which appears in all verb forms of compound tenses, which occupy a considerable number in the group of forms of a verb.

[4]Matthew Baerman, Scott Collier, Stem in a database of morphological complexity, 14th International Morphology Meeting, Budapest, Hungary, Workshop *Stems in inflections and lexeme formation*.

Starting from the final sound of the stem of the representative form of verbs (F1), as well as from the type and number of phonetic changes they undergo, we have compiled 25 representative models of formulas for the automatic generation of different verb forms in the Albanian language, which are illustrated with the verb *MËSOJ* (*learn*). According to this model there are automatically generated the forms of about 2990 other verbs.

Since the verb in the Albanian language has many forms, we are giving here only the models of how the formulas work for the present tense, simple past and perfect indicative tense, in all persons, singular and plural.

The verb *MËSOJ* (Conjugation I; class I; subclass 1; stem in *-o*) – **Indicative mood**						
	Simple Present	**Formulas**	**Simple Past**	**Formulas**	**Perfect**	**Formulas of present perfect**
unë	mësoj	F1+j	mësova	F1+va	kam mësuar	kam F6
ti	mëson	F1+n	mësove	F1+ve	ke mësuar	ke F6
ai/ajo	mëson	F1+n	mësoi	F1+i	ka mësuar	ka F6
ne	mësojmë	F1+jmë	mësuam	F5+m	kemi mësuar	kemi F6
ju	mësoni	F1+ni	mësuat	F5+t	keni mësuar	keni F6
ata/ato	mësojnë	F1+jnë	mësuan	F5+n	kanë mësuar	kanë F6

In the following list there are the formulas for the tenses of every mood, active and passive form, only in first singular person:

INDICATIVE MOOD

Present: *mësoj* (F1+j)
Imperfect: *mësoja* (F1+ja)
Simple past: *mësova* (F1+va)
Future: *do të mësoj* (do të F1+j)
Future II: *kam për të mësuar* (kam për të F6)
Future perfect: do të kem mësuar (do të kem F6)
Present nonactive: mësohem (F1+hem)
Imperfect nonactive: mësohesha (F1+hesha)
Simple past nonactive: u mësova (u F1+va)
Future nonactive: do të mësohem (do të F1+hem)
Future II: kam për t'u mësuar (kam për t'u F6)
Future perfect nonactive: do të jem mësuar (do të jem F6)
Present perfect: kam mësuar (kam F6)
Past perfect: kisha mësuar (kisha F6)
Pluperfect: pata mësuar (pata F6)
Future in the past: do të mësoja (do të F1+ja)
Future in the pat II: kisha për të mësuar (kisha për të F6)
Future perfect in the past: do të kisha mësuar (do të kisha F6)
Present Perfect nonactive: jam mësuar (jam F6)
Past Perfect nonactive: isha mësuar (isha F6)
Pluperfect nonactive: qeshë mësuar (qeshë F6)
Future past nonactive: do të mësohesha (do të F1+hesha)
Future past II: kisha për t'u mësuar (kisha për t'u F6)
Future perfect in the past nonactive: do të isha mësuar (do të isha F6)

SUBJUNCTIVE MOOD

Present: të mësoj (të F1+j)
Imperfect: të mësoja (të F1+ja)
Present nonactive: të mësohem (të F1+hem)
Imperfect nonactive: të mësohesha (të F1+hesha)
Present perfect: të kem mësuar (të kem F6)
Past perfect: të kisha mësuar (të kisha F6)
Present Perfect nonactive: të jem mësuar (të jem F6)
Past Perfect nonactive: të isha mësuar (të isha F6)

ADMIRATIVE MOOD

Present: mësuakam (F5+kam)
Imperfect: mësuakësha (F5+kësha)
Future: do të mësuakam (do të F5+kam)
Present nonactive: u mësuakam (u F5+kam)
Imperfect nonactive: u mësuakësha (u F5+kësha)
Future nonactive: do t'u mësuakam (do t'u F5+kam)
Present perfect: paskam mësuar (paskam F6)
Past perfect: paskësha mësuar (paskësha F6)
Present nonactive: qenkam mësuar (qenkam F6)
Imperfect nonactive: qenkësha mësuar (qenkësha F6)

SUBJUNCTIVE- ADMIRATIVE MOOD

Present: të mësuakam (të F5+kam)
Imperfect: të mësuakësha (të F5+kësha)
Present nonactive: t'u mësuakam (t'u F5+kam)
Imperfect nonactive: t'u mësuakësha (t'u F5+kësha)
Present perfect: të paskam mësuar (të paskam F6)
Past perfect: të paskësha mësuar (të paskësha F6)
Present Perfect nonactive: të qenkam mësuar (të qenkam F6)
Past Perfect nonactive: të qenkësha mësuar (të qenkësha F6)

CONDITIONAL MOOD

Present: do të mësoja (do të F1+ja)
Present II: kisha për të mësuar (kisha për të F6)
Present nonactive: do të mësohesha (do të F1+hesha)
Present II nonactive: kisha për t'u mësuar (kisha për t'u F6)
Present perfect: do të kisha mësuar (do të kisha F6)
Present Perfect nonactive: do të isha mësuar (do të isha F6)

OPTATIVE MOOD

Present: mësofsha (F1+fsha)
Present nonactive: u mësofsha (u F1+fsha)
Present perfect: paça mësuar (paça F6)
Present Perfect nonactive: qofsha mësuar (qofsha F6)

IMPERATIVE MOOD

Present: mëso (F1); mësoni (F1+ni)
Present nonactive: mësohu (F1+hu); mësohuni (F1+huni)

5. Conclusions

This algorithm has taken into account the verbs from the lexicon of Albanian vocabulary (https://gjuhashqipe.com/apps/fmgjsh).
Manual choice of the concrete verb conjugation is done for new verbs that are created or borrowed in Albanian, and then the algorithm performs the appropriate actions for the generation of all verb forms.
In case has been made an incorrect choice of the verb conjugation, the wrong forms are visible, and it only takes one more click to make the correct selection and the automatic generation of all verb forms. The results of this algorithm and research on the use of natural language inform us about the practical extent of morphological complexity for a language like Albanian and allow us to identify ways to improve the model.
Although we have worked on the morphological generation, this is very important even about the morphological analysis of Albanian language.
This software, as an application of the morphological structure of words, is closely interconnected to the software for Albanian spellchecker for MS Office (https://gjuhashqipe.com/softueret/drejtshkrimori). It is also of special importance in terms of preparing applications for tagging words in the corpus, for parsing, lemmatization, and in several natural language applications: text generation, machine translation, document retrieval, etc.
It can be used for genuine linguistic studies, but also for the acquisition of the language by the ordinary user, specifically the forms of words and their grammatical categories. Digital Morphology (https://gjuhashqipe.com/apps/kulla) is useful for both students and teachers of the Albanian language, because using this software students can test, evaluate, and improve their knowledge, while teachers can use it to perform practical tasks with examples from the vocabulary of the Albanian language, making learning the morphology of Albanian even more attractive and interactive.

6. References

[1] Bolshakov, I. A.; Gelbukh, A.: Computational Linguistics Models, Resources, Applications, 2004, 186 pp; ISBN 970-36-0147-2, www.gelbukh.com/clbook
[2] Çepani, A.; Çerpja, A.: Hyrje në gjuhësinë kompjuterike (tekst universitar), Fakulteti i Historisë dhe i Filologjisë, “Albas”, Tiranë, ISBN 978-9928-02-833-4, 2017, 232 f.
[3] *Gjuha letrare shqipe për të gjithë. Elemente të normës letrare kombëtare*. Kostallari A.; Lafe E.; Totoni, M.; Cikuli, N. ShBLSh. Tiranë, 1976, 294 f.
[4] *Gramatika e gjuhës shqipe*, I, II, Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, Tiranë, 2002.
[5] Jurafsky, D.; Martin, J. H: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, 2000; see www.cs.colorado.edu/~martin/slp.html.
[6] Shishani, L., Çerpja, A.: “Gjuha shqipe dhe programi për drejtshkrim AS 2.0”, Gjuha jonë, n. 1-4, 2005, f. 126-134.