Morphological parsing of Albanian language: a different approach to Albanian verbs

Ilir Çollaku
Istanbul Technical University, ilir.collaku@itu.edu.tr

Eşref Adali
Istanbul Technical University, adali@itu.edu.tr

Morphological parsing of Albanian language: a different approach to Albanian verbs

Ilir Çollaku[1], Eşref Adalı[2]

Faculty of Computer Engineering and Informatics Istanbul Technical University,
ilir.collaku@itu.edu.tr[1], adali@itu.edu.tr[2]

Abstract. The very first step when processing a natural language is creating a morphological parser. Verbs in Albanian language are the most complex area of inflection. Besides irregular verbs, the ways in which the regular verbs change their form while being inflected are hardly definable, and the number of exceptions is huge.
In this paper, a different approach to Albanian verbs is made. Unlike traditional classification, based on the inflection themes they take, verbs are classified into different verb groups. This way, the inflection process looks clearer and more regular, as the affix remains the only changeable part of the inflected verb. This way of approach, makes us able to process the Albanian verbs simpler and easier.

Keywords: NLP Albanian, morphological parsing, tagging, Albanian verbs, inflection theme

1. Introduction

Verbs in Albanian language are definitely the most complex among all parts of speech. Because of their very rich entity of inflection forms, they are considered as the widest class of words and the hardest one to fully master. Inflection forms in Albanian are built with changes in verb form and by adding suffices, whereas the descent relationship between them is very complicated [1]. Verb inflection forms, together with the construction of plural forms of nouns, are considered as two major difficulties in Albanian morphology and its processing. Besides the political situation reigned in Albania and Albanian-populated lands during last century, and the fact that no standard Albanian language existed until the late 1970's, the above mentioned difficulties can also be considered as a reason why there is a lack of studies and research on NLP (natural language processing) Albanian. Thus, Albanian language is considered as one of the most challenging languages to process [2].

2. Verbs in Albanian

Verbs in Albanian have their variable and constant forms. Variable forms of Albanian verbs have six moods: indicative, subjunctive, conditional, admirative, optative and imperative; nine tenses: present, imperfect, past simple, past, past perfect, (past) past perfect, future, future past and future front; and two voices: active and passive. Constant forms of verbs are: past participle, infinitive, present participle and negative. Regarding all these attributes, we can say that a single verb falls in 47 different inflectional forms1. In almost every book of Albanian grammar, verbs are grouped based on their inflection. According to traditional grouping, there are three groups of verbs: first inflection verbs (verbs ending with -j in their singular 1st person indicative present tense active voice form, e.g. punoj 'I work'), second inflection verbs (those ending in consonants, e.g. shoh 'I see'), and third inflection verbs (those ending in vowels, e.g. ha 'I eat').

2.1 Traditional grouping – a short analysis

First inflection verbs indeed have similar attributes. However, their inflections sometimes can differ, e.g.

la.j - la.rë ‘I wash - to wash/washed’, but
mëso.j - mësua.r ‘I learn - to learn/learnt’

(note that the inflection theme of the second verb changes in its past participle, while that of the first one remains same), and this can be useful in defining inflection rules when processing the language. Second and third inflection verbs, on the other hand, differ from one another so much in inflection, such that sometimes no similarities can be determined between verbs of the same group, e.g. the verb pres ‘I wait’ changes its theme in past simple to prit.a ‘I waited’, while tremb, ‘I scare’ doesn’t change it, tremb.a ‘I scared’, and no differences can be determined between verbs of different groups. A second inflection verb inflects same in imperfect tense as one of the third inflection, e.g.

ha - ha.ja ‘I eat - I was eating’ and
hap - hap.ja ‘I open - I was opening’

Thus, when talking about morphological processing of Albanian verbs, in order to be able to define inflection rules, they must be grouped in a more specific way.

2.2 A different grouping of Albanian verbs

A different method, but not unfamiliar to Albanian language, is grouping verbs based on their inflection themes. When talking about verb theme, everyone thinks of its representative form, the form it appears in dictionaries - in Albanian language it is the first person singular present tense indicative mood active voice form of the verb (e.g.: punoj ‘I work’, laj ‘I wash’, ha ‘I eat’, marr ‘I take’), but in morphology verb theme has another mean: it is the stem of the verb from which inflection forms are built, and is called the inflection theme of the verb.

1constant forms have been included, imperative mood has only the 3rd person, not all tenses are applicable to all moods, and not every verb has its active and passive voice
For example, an inflection theme of the verb punoj ‘I work’ is puno., from which a myriad of forms are built, like:

puno.j ‘I work’
puno.n ‘you(sg) work/he,she works’
puno.jmë ‘we work’ puno.ni ‘you(pl) work’ puno.jnë ‘they work’
puno.ja ‘I was working’ puno.va ‘I worked’
...

The same verb comes also with a different theme, that is punua.:

punua.r ‘working’
punua.m ‘we worked’
punua.kam ‘I have been working (admirative)’
punua.kësha ‘I had been working (admirative)’
...

Verbs with a single inflection theme aren’t few, e.g.:

fshi. ‘I clean’ la.j ‘I wash’ pi. ‘I drink’

But most of Albanian verbs have two or more inflection themes, and some of them can have up to eight [1].

mëso.j ‘I learn’
mësua.r ‘to learn/learnt’

blua.j ‘I grind’ blo.va ‘I ground’ blu.het ‘it is ground’

shoh. ‘I see’
sheh. ‘you(sg) see’ shih.ni ‘you(pl) see’ pa.shë ‘I saw’

dua.n ‘they want’ do.ja ‘I was wanting’ desh.a ‘I wanted’
dash.ur ‘to want/wanted’ du.hem ‘I am wanted’

the.m ‘I say’
thua. ‘you(sg) say’ tho.të ‘he/she says’ thu.het ‘it is said’
tha.shë ‘I said’
thë.në ‘to say/said’
thën.çi ‘you(pl) say (optative)’

Fig. 1. Examples of verbs with two, three, four, five, and seven inflection themes. Based on the last vowel of their inflection theme, the number of inflection themes, suffices they take and other structural changes while being inflected, verbs in Albanian language are grouped in two major classes: verbs inflection themes of which end with a vowel or a heap of vowels, and verbs inflection themes of which end with a consonant. Further, depending on the number of inflection themes, each class is separated to groups of verbs with one, two, three, up to eight inflection themes [1].

3. Parsing Albanian verbs – the solution proposed

This new method of grouping verbs based on their inflection themes, also changes the approach to verbs in computational processing. Since Albanian language has extremely rich inflectional paradigms and principles they follow aren’t systematic enough, they are not appropriate to be presented by algorithms. Even if the necessary number of algorithms were designed for groups of similar verbs, the number of exceptions would remain enormous.

3.1 Parsing based on the inflection themes

The number of verbs with similar structure but different inflectional specifications in Albanian language isn’t small, e.g. verbs luaj ‘I play’ and bluaj ‘I grind’ are very similar, but the way they inflect is different:

lua.j - luajt.a - luajt.ur ‘I play - I played - to play/played’, while

blua.j - blo.va - blua.r ‘I grind - I ground - to grind/ground’

Thus, the proposed solution which could make parsing Albanian verbs easier, is putting all their inflection themes to the corpus. For example, auxiliary verbs kam ‘I have’ and jam ‘I am’ have the greatest number of inflection themes, the first one has seven:
ka.m ‘I have’
ke.ni ‘you(pl) have’
kish.a ‘I had’
pat.a ‘I have had’
paç.im ‘we have (optative)’
pas.kam ‘I’ve had (admirative)’
ki.ni ‘you(pl) have (imperative)’

The other has eight:

ja.m ‘I am’ je.mi ‘we are’ është. ‘he,she is’ ish.a ‘I was’
qe.shë ‘I have been’
qen.kam ‘I have been (admirative)’ qo.fsh ‘you(sg) be (optative)’ ji.ni ‘you(pl) are (imperative)’

In this case, for two verbs kam and jam, all 15 records; ka, ke, kish, pat, paç, pas, ki, and ja, je, është, ish, qe, qen, qo, ji should be put to the corpus. Thinking in similar way, plural forms of some nouns should also be kept in the corpus (e.g. palë - palë 'pair - pairs', kalë - kuaj 'horse - horses', djalë - djem 'boy - boys'). As it can be seen from the above example, the three nouns are very similar to each other (they all end up with '-alë'), but they construct their plural forms in completely different ways.

3.2 Examples of verb inflection

Placing verb inflection themes on the corpus makes the definition of morphologic rules more likely in computational environment. In contrast to inflection themes, suffices that the verbs take while being inflected are more stable. They do not change at all, but only in certain cases, when depending on the theme's ending, an additional phonem (-j- or - n-) is placed between the theme and suffix. Suffices of imperfect tense are constant for all verbs, they are added to the inflection form of the verb to build its imperfect tense indicative and admirative forms, for example:

verb kam 'I have' (inflection themes: ka., ke., kish., pat., paç., pas., ki.) Indicative form Admirative form
'I was having' 'I have been having'

kish.a pas.kësh.a kish.e pas.kësh.e kish.te
pas.kësh.- kish.im pas.kësh.im kish.it pas.kësh.it
kish.in pas.kësh.in

verb jam 'I am' (infl. themes: ja., je., është., ish., qe., qen., qo., ji.) Indicative form Admirative form
'I was being' 'I have been being'

ish.a qen.kësh.a ish.e qen.kësh.e ish.te
qen.kësh.- ish.im qen.kësh.im ish.it qen.kësh.it
ish.in qen.kësh.in

verb mësoj 'I learn' (inflection themes: mëso., mësua.) Indicative form Admirative form
'I was learning' 'I have been learning'

mëso.(j)a mësua.kësh.a mëso.(j)e mësua.kësh.e mëso.(n)te
mësua.kësh.- mëso.(n)im mësua.kësh.im mëso.(n)it
mësua.kësh.it mëso.(n)in mësua.kësh.in

verb dal 'I go out' (inflection themes: dal., del., dil., dol.) Indicative form Admirative form
'I was going out' 'I have been going out'
dil.(j)a dal.kësh.a dil.(j)e dal.kësh.e dil.te
dal.kësh.- dil.(n)im dal.kësh.im dil.(n)it dal.kësh.it
dil.(n)in dal.kësh.in

Fig. 2. Examples of verb inflection in imperfect tense, indicative and admirative forms.

Conclusion

Including inflection themes to the corpus enables us to handle inflection process of verbs more systematically. Since in Albanian language the suffix is the part which keeps the information about verb attributes (like mood, tense and voice), making verb themes static, reduces the scope of changes in verb structure and eases inflection rules definition. Besides this, it also avoids definition of extra inflection rules based on which the themes are built, and which, because of a lot of exceptions, still would not offer a complete solution.
The only disadvantage here is keeping more than one record in the corpus for verbs which come with several inflection themes, instead of keeping one record per verb. But again, this remains a better solution than creating dozens of rules for each group of verbs, which also makes the processing last longer.

References

1. Memushaj R., “Shqipja Standarde - Si ta flasim dhe ta shkruajmë”, Botimet Toena 2004, pg.87-94
2. Trommer J., Kallulli D., “A Morphological Tagger for Standard Albanian”
3. Piton O., Lagji K.,”Morphological study of Albanian words, and processing with NooJ”, 2007