Modeling a Two-Level Formalism for Inflection of Nouns and Verbs in Albanian

Arbana Kadriu
South East European University
Macedonia

1. Introduction

The core task of computational morphology is to take a word as input and produce a morphonological analysis for it. Morphotactics defines the model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside of a word (Jurafsky & Martin, 2000). But there are situations where the word formation process is not just joining of morphemes, such as assimilation, insertion, duplication, etc., and this are the situations where the phonological rules show up. Phonological rules may apply and change the shape of morphs (Mitkov, 2003).
Many linguists have modeled phonological rules, but it is considered that the most successful one is the model called *two-level morphology* (Koskenniemi, 1983). The two-level morphology model has been proved successful for formalizing the morphology of very different languages (English, German, Swedish, French, Spanish, Danish, Norwegian, Finnish, Russian, Turkish, Arab, Aymara, Swahili etc.) (Uibo, 2003). This system is used even for conversion between different writing systems (Maleki & Ahrenberg, 2008).
Thus, we can expect that the model is in fact universal and it may be possible to describe Albanian morphology in this framework as well.
Extensive research is done in this area in widely-used languages. For the Albanian language, to our knowledge, there is no such research that offers two-level formalism for any grammatical category. Initial point of the research presented here is a list of 4200 Albanian verbs in ten different tenses and a list of more than 100 Albanian nouns, where each record of the list holds the noun's form in all possible cases. This input data is used to obtain suffixes and rules that define the phonological alternations during concatenations.

2. Verbs in Albanian

Verbs are the most complex area of Albanian inflection (Trommer, 1997). An Albanian verb can be in one of the 6 possible moods: *indicative, admirative, conjunctive, conditional, optative,* and *imperative*.
The *indicative mood* is used for simple statements, declarations, etc., such as **shkruaj** (*I write*).
The *admirative mood* is used to make statements of admirations, especially when surprised or in unexpected situations. *Conditional mood* and *subjunctive mood* are used to express

possibility. *Optative mood* is used for wishes or for curses. The *imperative* is used for orders, commands, or demands.
Verbs are conjugated in six tenses: *present, imperfect, future, past, present perfect, past perfect*. Each mood has *tenses*, and each tense has 6 *persons*: 3 for singular and 3 for plural. Only indicative mood has all six tenses.
Verbs are listed in the vocabulary in the present first person singular of the indicative: **unë shkoj** *(I go)* - **shkoj** *(to go)*. Some other examples are: **shkruaj** *(to write),* **shkoj** *(to go),* **ndihmoj** *(to help),* **pres** *(to wait),* **përsëris** *(to repeat),* **mësoj** *(to learn) etc.*
Verbs in Albanian do not have infinitive. There is a form, so called *paskajore* (translated as *infinitive*), that assumes many functions of the infinitive. Besides this, Albanian verbs have the *participle* and the *gerund*. For the verb **shkoj** (*to go*), participle is **shkuar**, gerund is **duke shkuar**, and infinitive **për të shkuar**.
Table 1 shows the conjugation of the verb **notoj** (*to swim*) in indicative mood, in the first four tenses. The future tense often is formed with the present subjunctive and the particle *do*. Table 2 shows the conjugation of this in indicative mood, in perfect tenses. The present perfect tense and the past perfect tense are formed with the participle form and finite forms of the auxiliaries ***kam*** *(to have*), or ***jam*** (*to be*).

	Present	**Imperfect**	**Past**	**Future**
unë (I)	notoj	notoja	notova	do të notoj
ti (you)	noton	notoje	notove	do të notosh
ai, ajo (he, she)	noton	notonte	notoi	do të notojë
ne (we)	notojmë	notonim	notuam	do të notojmë
ju (you)	notoni	notonit	notuat	do të notoni
ata (they)	notojnë	notonin	notuan	do të notojnë

Table 1. The conjugation of the verb **notoj** (*to swim*) in indicative mood – present, imperfect, past, future

	Present Perfect	**Past Perfect**
unë (I)	kam notuar	kisha notuar
ti (you)	ke notuar	kishe notuar
ai, ajo (he, she)	ka notuar	kishte notuar
ne (we)	kemi notuar	kishim notuar
ju (you)	keni notuar	kishit notuar
ata (they)	kanë notuar	kishin notuar

Table 2. The conjugation of the verb **notoj** (*to swim*) in indicative mood - the perfect tenses

Like in English, verbs in Albanian may be *transitive* or *intransitive*, of *active* or *passive* voice. The intransitive verbs (including reflexive verbs) are called in Albanian grammar *non-active*.

3. Nouns in Albanian

Albanian nouns are inflected by gender (masculine, feminine and neuter) and number (singular and plural). There are 5 declensions with 5 cases (nominative, accusative, genitive,

dative and ablative). The cases apply to both definite and indefinite article. The equivalent of a genitive is formed by using the prepositions *i/e/të/së*. The equivalent of an ablative is formed by using the prepositions *prej*. It should be mentioned that inflection of the Albanian nouns is realized through suffixes and no prefixes are used.
Table 3 shows the declension of the masculine noun *lis* (type of tree). Table 4 shows the declension of the feminine noun *fushë* (field).
The base form is considered the indefinite article nominative (Agalliu et al., 2002). The definite article can be in the form of noun suffixes, which vary with gender, case and number. For example, in singular nominative, masculine nouns often add *-i* or *-u*:

- *lis* (a tree) / *lisi* (the tree);
- *mik* (a friend) / *miku* (the friend).

	Indef. Sing.	**Indef. Pl.**	**Def. Sing.**	**Def. Pl.**
Nom.	lis (tree)	lisa (trees)	lisi (the tree)	lisat (the trees)
Gen.	lisi	lisave	lisit	lisave
Dat.	lisi	lisave	lisit	lisave
Accu.	lis	lisa	lisin	lisat
Abl.	lisi	lisash	lisit	lisave

Table 3. The declension of the masculine noun *lis* (tree)

	Indef. Sing.	**Indef. Pl.**	**Def. Sing.**	**Def. Pl.**
Nom.	fushë (field)	fusha (fields)	fusha (the field)	fushat (the fields)
Gen.	fushe	fushave	fushës	fushave
Dat.	fushe	fushave	fushës	fushave
Accu.	fushë	fusha	fushën	fushat
Abl.	fushe	fushave	fushës	fushave

Table 4. The declension of the feminine noun *fushë* (field)

4. Two-Level Morphology

This model includes two components:

- the phonological rules described through finite state transducers,
- the lexicon, which includes the lexical units and the morphotactics.

The formalism called *two-level phonology* is used in the two-level morphology model. The Koskenniemi's model is two-level in the sense that a word is described as a direct, symbol-to-symbol correspondence between its lexical and surface form (Antworth, 1990).
The two-level rules consist of three components: the correspondence, the operator and the environment.
Every pair lexical symbol-surface symbol is called correspondence pair. The notation for this correspondence is: *lexical symbol : surface symbol*.
For the first character of the previous example, we would write m:m, while for morpheme boundary we have +:0.
The operator defines the relation between the correspondence and the environment, where it appears. There are four operators, illustrated as follows:

- => - the correspondence appears only in this environment, but not always

- <= - the correspondence appears always in this environment, but not only in this one
- <=> - the correspondence appears always and only in this environment
- /<= - the correspondence never appears in this environment

The third component relates to the environment and specifies the phonological context where a specific phenomenon happens. The notation for this component is realized through the underline sign "_", called environment line, and its general form is LC__RC, where LC denotes the left context, while RC denotes the right context.

4.1 Automatic learning of morphotactics and two-level rules

Theron&Cloete used an augmented version of edit distance to align an underlying and surface string and discover morpheme boundaries (Theron & Cloete, 1997). They then look for insertions, deletions, and replacements in the alignment to find the location of a two-level rule, and look for the minimal context surrounding the rule using extensions of the heuristics in Johnson (Johnson, 1984) and Touretzky (Touretzky et al., 1990).

For automatic acquisition of two-level phonological rules for the inflection of the Albanian verbs and nouns, this model is adapted and upgraded.

They use the notion of string edit sequences assuming that only insertions and deletions are applied to a root form to get the inflected form. They determine the root form associated with an inflected form and consequently the suffixes and prefixes by exhaustively matching against all root words (Oflazer & Nirenburg, 1999).

As a first step for this research, a set of pairs base_noun (base_verb) - inflected_noun (inflected_verb) is constructed. The output of this phase is used as input for the second phase of the algorithm. For every pair of this input, according to the Theron&Cloete algorithm, a string edit sequence is constructed, using the insert, delete and replace operations.

With the aim to improve the results, a few string edit distance algorithms are tested and the best results are gained with the Brew string distance (Kadriu & Zdravkova, 2008).

After all the transformations, for all input pairs that are processed, the next step is to deal only with the special pairs. A special pair is every pair that presents deletion, insertion or replacement of a character.

For every special pair, the context in which they appear is constructed as follows: first left neighbour, then the first right neighbour, followed by the second left neighbour, then second right neighbour, and so on. At the end of this sequence a special pair is written, called *marked pair*. Some other special signs are also used, depicting the start of the string (SOS), the end of the string (EOS), and the sign (OOB), which is used when one context (left or right) is longer than the other one.

Using the resulting sequences, a minimal acyclic finite automaton is constructed, and it has only one start and one final state. The edges of this automaton represent the special pairs, while all terminal edges represent the marked pairs. The automata are constructed using xfst tool (Beesly & Karttunen, 2003), considering the constructed contexts as regular expressions.

If for a marked pair all paths go through some shortened path, the new found path is considered as context for that pair.

The next step is for every marked pair *x:y* to answer the question: 1) is this context the only one in which this pair occurs, and 2) is this pair always occurring in this context.

To answer the first question, all the paths that contain the marked pair should be passed. If they all have a common segment, the answer is yes. In other words, this means that for this context the rule => is true.
For the second question, all the terminal edges that have left component x are verified. If they all have right component equivalent to y, the answer for the second question is yes. This means that for this context the rule <= is true.
If the answer is positive for both questions, that means that this marked pair occurs always and only in this environment, i. e. for this context the rule <=> is true.
At the end, from the set of gained rules, the rule <=> with the shortest context is taken. If there is no <=> rule, the shortest context for the <= rule and/or => rule is picked up. If both have a common context, then they are concatenated in a single <=> rule.

5. Acquisition of morphotactics and two-level rules

5.1 The achieved morphotactics for verbs inflection

As it was mentioned, perfect tenses are formed using the participle form and finite forms of the auxiliary verbs. Since the participle is fixed and doesn't suffer phonological changes, these tenses are not covered by the research presented in this paper. Imperative mood has only the present tense and is conjugated only in the second person – singular and plural. For this reason, this mood is also not considered for further processing in this paper.
The input for the automatic acquisition of morphotactics and two-level phonological rules is a list of 70 verbs in ten different tenses: indicative (present, imperfect, past, future), conjunctive (present, past), admirative (present, imperfect), conditional (present), optative (present). Table 5 shows all 60 input forms used in our system for the verb **ftoj** (*to invite*).

Indicative				**Conjunctive**	
Present	*Imperfect*	*Past*	*Future*	*Present*	*Past*
ftoj	ftoja	ftova	do të ftoj	të ftoj	të ftoja
fton	ftoje	ftove	do të ftosh	të ftosh	të ftoje
fton	ftonte	ftoi	do të ftojë	të ftojë	të ftonte
ftojmë	ftonim	ftuam	do të ftojmë	të ftojmë	të ftonim
ftoni	ftonit	ftuat	do të ftoni	të ftoni	të ftonit
ftojnë	ftonin	ftuan	do të ftojnë	të ftojnë	të ftonin
Admirative		**Conditional**	**Optative**		
Present	*Imperfect*	*Present*	*Present*		
ftuakam	ftuakësha	do të ftoja	ftofsha		
ftuake	ftuakëshe	do të ftoje	ftofsh		
ftuaka	ftuakësh	do të ftonte	ftoftë		
ftuakemi	ftuakëshim	do të ftonim	ftofshim		
ftuakeni	ftuakëshit	do të ftonit	ftofshi		
ftuakan	ftuakëshin	do të ftonin	ftofshin		

Table 5. All input forms of the verb **ftoj** (*to invite*)

In view of the fact that each tense has six persons, it indicates that we have 4200 different forms of verbs as input for our system. Some of the tenses are formed with particles such as *do, u, të*. They are fixed and do not change. For this reason, we are interested only on the verb part and will not consider the particles during the morphological analysis of the verbs in those tenses. No particular distinction between transitive and intransitive verbs is made here.

As it was mentioned, verbs are listed in vocabularies in the present first person singular of the indicative. Given this fact, we use this form as a base form for further processing. All other forms as considered as derived forms. So, for each tense and for each verb, a list of 6 pairs base_verb-derived_verb is formed. For the verb **ftoj** (Table 3), imperfect tense, we have the following list of six pairs as an input.

ftoj ftoja
ftoj ftoje
ftoj ftonte
ftoj ftonim
ftoj ftonit
ftoj ftonin

For each tense, there is an input list of 420 such pairs of verbs. The output of the first step of the algorithm is a list of pairs *base_verb+obtained_suffix - derived form*. For the above input, the automatically achieved output is as follows:

ftoj+a ftoja
ftoj+e ftoje
ftoj+im ftonim
ftoj+in ftonin
ftoj+it ftonit
ftoj+te ftonte

So, from this process, the suffixes used for inflection in every particular tense are obtained. Table 6 shows all suffixes obtained.

Mood	**Tense**	**Gained suffixes**
Indicative	Present	më, në, m, n, t, i, ni, sh, im, in, o, 0
	Imperfect	a, e, it, te, *ësh,* sha, she, *shim, shin, shit,* j, ja, je
	Past	ëm, ët, ën, va, ve, am, at, an, u, të, ë, m, n, sh, më, në, im, in, ni, a, e, i, *ësh,* 0
	Future	më, ni, në, sh, ë, im, in, i, t, n, 0
Conjunctive	Present	më, ni, në, sh, in, im, i, t, n, ë, *ësh,*0
	Past	a, e, it, im, in, te, *ësh,sha, she,* ja, je, nim, nin, nit
Admirative	Present	*kam, kan,* ka, ke, *kemi, keni*
	Imperfect	kësh, kësha, këshe, këshim, këshin, këshit
Conditional	Present	a, e, it, im, in, te, *sha, she,* ja, je, j
Optative	Present	im, in, sha, shi, sh, të, shim, *shin*

Table 6. Suffixes obtained for every particular tense

It should be mentioned that there are some cases where it is intervened manually (suffixes in italic):

- There are some letters in Albanian that consist of two characters. One such case is the letter **sh**. There were cases where instead of suffixes *sha, she, shim, shin, shit*, the suffixes ha, he, im, in, it were obtained (and the insertion of character s was suggested).
- The suffixes of the admirative mood, except for present tense suffixes *ka* and *ke*, come out without the first character *k*. For example, instead of the suffix *këshim,* the suffix *ëshim* is obtained.

The suffix *ësh* does not appear as a suffix in itself. Instead, it appear as insertion of character *ë* plus suffix *sh*.

5.2 The achieved morphotactics for nouns inflection

List of over 100 Albanian nouns in all cases is used to automatically learn the morphotactics and two-level phonological rules. As base word is considered the indefinite singular of nominative. The following tags are used to describe special cases: S - stays for singular, P – plural, indef – indefinite, def – definite, nom – nominative, gen – genitive, dat – dative, acc – accusative, abl – ablative.

For example, the noun *ditë* (*a day*), corresponds to this input record:

- *ditë (S_indef_nom&acc)*
- *dite (S_indef_gen&dat&rrj)*
- *dita (S_def_nom)*
- *ditën (S_def_acc)*
- *ditës (S_def_gen&dat&rrj)*
- *ditë (P_indef_nom&acc)*
- *ditëve (P_indef-gen&d_def_gen&dat&abl)*
- *ditët (P_def_nom&acc)*
- *ditësh (P_indef_abl)*

From the above list, it can be noticed that the same form can be used for several cases. This is so because, as explained in the second part of this paper, same cases are defined using prepositions, and other cases are defined from the context.

The procedure is applied to all 8 cases in separate (the first case is used as a base form). After applying the first step of the above described algorithm, we obtain all proposed (by the system) segmentations, which are in a form as the one shown below:

- *bukë+e buke* (bread)
- *burrë+i burri* (man)
- *dashuri+e dashurie* (love)
- *dhembje+je dhembjeje* (pain)
- *dhomë+e dhome* (room)
- *diell+i dielli* (sun)

Following inflectional suffixes are automatically obtained:

Case	**Gained suffixes**
S_indef_gen&dat&abl	e, je, i, u, ri
S_def_nom	i, u, a, ja, 0
S_def_gen&dat&abl	t, së, s, it, rit, ut
S_def_acc	n, në, in, un
P_indef_nom&acc	a, e, em, ëz, j, ë, arë, ra, 0
P_indef-gen&d_def_gen&dat&abl	ve, ave
P_def_nom&acc	të, ët, at, et, t
P_indef_abl	sh

The learned suffixes are implemented in the lexicon used for defining two-level model of Albanian nouns.

5.3 The achieved phonological rules

The segmentations achieved in the previous phase are used as input for the second phase, where the morpho-phonological alternations are learned. Taking into consideration the fact that we want to create a model that will include all verbs/nouns, and not distinct models for each tense/case, we automatically reduce the rules so that they will not conflict with each other.

For each special pair, merge left-arrow rules with right-arrow rules into a double-arrow rule with intersecting context. If several contexts are available for some rule, the new rule will have the intersected context. If for some special pair same-arrow rules have several contexts with an empty intersection, make a new rule with disjunctive contexts. Finally, resolve conflicts as explained in any two-level literature.

For example, for all cases we got the special pair *ë:0*, but in different context:

- *ë:0* <=> _ +:0
- *ë:0* <=> _ +:0 i:i t:t EOS
- *ë:0* => _ +:0 i:i n:n EOS
- *ë:0* => _ +:0
- *ë:0* => _ +:0
- *ë:0* => _ +:0
- *ë:0* => _ +:0
- *ë:0* => _ +:0

After the reduction, we got only a single rule for the above obtained rules: *ë:0* =>_ +:0.

The automatically learned rules were used as a base for further extension, testing them and manually improving the "holes" in the system:

– The rules that involve insertion of characters *k, s, ë* or replacement of some characters to *k, s, ë* are removed (for reasons explained in the previous section).

– The rules with a too long context are removed, as it is for example the following rule:

0:e => SOS r:r ë:ë n:n k:k ë:ë s:s h:h +:0 +:0 _ EOS OOB OOB OOB OOB OOB OOB OOB OOB1

– For some of the marked pairs the context is too short to describe the environment where they appear. For example, rule "*0:n => +:0 _* " is extended to the rule "*0:n => +:0 _ t:t e:e*".

[1] EOS, OOB, SOS are special signs used in the algorithm for machine learning of rules

– Some special pairs appear in the context of some rules, but they never appear as marked pair. For example, in the rule "*o:u* => _ *0:a +:0*" we have the special pair *0:a*, but this never appears on the left side of any rule (as marked pair).

After improving these gaps, there are twelve two-level rules implemented for the verb inflection:

1. 0:o => _ n:n i:i #:#
2. u:0 => _ a:0 +:0
3. a:0 => u:0 _ +:0
4. m:0 => [e:e | e:0] _ +:0
5. j:0 => _ +:0
6. 0:n => +:0 _ t:t e:e
7. t:s <=> _ +:0 t:t e:e #:#
8. h:0 => [o:o | o:u] _ e:0 m:0
9. e:0 => _ [m:0 +:0 | +:0]
10. o:u => _ [h:0 e:0 m:0 | j:0]
11. 0:f => +:0 _ [s:s h:h | t:t ë:ë]
12. 0:a <=> o:u [h:0 e:0 m:0 | j:0] +:0 _ k:k

For the noun inflection, the following rules were implemented:

1. *ë*:0 => _ +:0 | _ r:r +:0
2. e:0 => _ +:0
3. 0:[a | e | ë] => _ +:0 v:v e:e
4. u:0 => _ a:a +:0
5. 0:r => +:0 _

6. Evaluation of the implemented system

The Albanian alphabet has 36 letters: a, e, i, o, u, ë, b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, x, y, z, ç, dh, gj, ll, nj, rr, sh, th, xh, zh. The last nine letters are concatenation of two characters. Since there are no phonological alternations obtained for these letters, they are dropped out when defining the alphabet used in the system – when they show up in the words, they are considered as distinct characters.

The system is implemented using PC-KIMMO environment (Antworth, 1995). It includes a lexicon file, a rules file and a file that contains list of verbs in their base form – the present first person singular of the indicative. The lexicon file includes the list of suffixes and the defined morphotactics for creating inflectional form of verbs and nouns. The rules file consists of several finite state automata that describe the defined rules.

6.1 Verbs system

An aspect that we considered while implementing the rules is the fact that there are cases where verbs ending in *em*, follow two segmentations, where one is semantically incorrect. For example, for the verb **ankohem** (*to complain*) both a semantically correct (1) and a semantically not correct (2) segmentation will be produced:

1. ankohem+ [V(ankohem)+AGENT-zero]
2. ankohem+m [V(ankohem)+AGENT-m]

The same situation is with verbs ending in *j*, when adding the suffix *j* or suffix that starts with the character *j*. That is why we added two disallowing rules, that is, rules that define that these correspondences never occur in this context:

1. m:0 /<= e:e _ +:0 m:m #:#
2. j:0 /<= _ +:0 j:j

After all the phonological and morphotactical rules are implemented, the system is tested on 4200 different forms of verbs. The result obtained is that all verbs are correctly segmented for the indicative present, indicative imperfect, indicative past. For the other tenses, in total 84 verbs are not segmented. Twelve of them (two verbs in six different cases) are result of situations where the rules with a too long context are removed.
In Albanian, there are cases of so-called irregular verbs, or verbs that in conjunctive past, admirative (present and imperfect), conditional and optative are derived from the participle and not from the present first person singular in the indicative mood. Since we used the present first person as a base form in input data, these occurrences are not modeled in our system. This case would be a subject of further research. Even with this "weakness" of the system, it has successfully segmented 98% of the verbs.

6.2 Nouns system

An aspect that we considered while implementing the rules is the fact that some gained rules produced morphologically correct, but semantically non-correct segmentations. It is the case when a suffix that corresponds to a masculine noun adds to a feminine noun. For example, if we consider the noun *britaniku* (*the Britain* – masculine). The system produces the semantically correct segmentation – *britanik+u* (a Britain masculine + suffix u), but also the semantically incorrect segmentation – *britanike+u* (a Britain feminine + suffix u). This is the case with some feminine nouns that end in *e* or *ë*. That is why we added two disallowing rules, that is, rules that define that these correspondences never occur in this context.

- *ë:0 /<= _ +:0 ë:ë [s:s | n:n | t:t | #:#]*
- *e:0 /<= _ +:0 [u:u | i:i | e:e | ë:ë]*

After all the phonological and morphotactical rules are implemented, the system is tested on 856 nouns that were not in their base form. These are seen verbs (verbs that were used for training). Table 7 gives a picture for the "negative" results when testing the system. All others are correctly segmented and produced only a single segmentation.

Case	No segmentation	Two segmentations
S_indef_gen&dat&abl	0	**3**
S_def_nom	0	5
S_def_gen&dat&abl	2	0
S_def_acc	2	0
P_indef_nom&acc	8	4
P_indef-gen&d_def_gen&dat&abl	8	0
P_def_nom&acc	8	0
P_indef_abl	8	0

Table 7. All "error" occurrences

All cases that contain two segmentations are situations when the fifth rule is applied (the insertion of character *r*).
For example, for the noun *syri* (*the eye*), we get the segmentation *sy+ri* (the suffix *ri* is added), but also the segmentation *sy+i* (here the character r is inserted).
The gained no-segmentation situations for cases: S_def_gen&dat&abl and S_def_acc are as a result of the fact that suffixes *-ës* and *–ën* are not obtained automatically from the first part of the algorithm. These suffixes are added to the lexicon and after that we did not have any no-segmentation situation for these two cases. Thus, we are left with only eight non-segmented nouns in four cases in plural. These are situations with so-called irregular nouns or nouns that undergo more complicated phonological alternations. In fact, as it was mentioned in section 6, rules are also produced for these alternations, but they were not considered for implementation, having in mind that a longer input list of "irregular" nouns would produce better results. Below is the list of pairs *S_indef_nom - P_indef_nom* of all those nouns:

- *babë - baballarë (a father – fathers)*
- *diell - diej (a sun – suns)*
- *djalë - djem (a boy – boys)*
- *dorë - duar (a hand – hands)*
- *natë - net (a night – nights)*
- *njeri - njerëz (a person – persons)*
- *vëlla - vëllezër (a brother – brothers)*
- *vit - vjet (a year – years)*

It should be mentioned that the other three cases are similar to the case *P_indef_nom*. It means that if the rules will be implemented for one of the cases, it will work for all other three cases.
The next step was testing of the system on unseen nouns. We used 416 nouns, extracted from a tagged text – the initial part of a novel (Trommer & Kallulli, 2004). To our knowledge, this is the only tagged text in Albanian. From these nouns, 405 nouns were correctly segmented. The remaining 9 nouns were irregular nouns, mentioned above (on seen nouns).

- *duar (hands) – twice*
- *miqve (to the friends)*
- *motrës (to the sister)*
- *njerëz (persons) – three times*
- *njerëzit (the persons)*
- *njerëzve (to the persons)*

This means that the system successfully segmented 98% of unseen nouns.

7. Conclusions

The research presented here is of a great significance as an already known methodology has been extended and put in another linguistic framework - Albanian verbs and nouns. It also describes a methodology for defining two-level formalism for a specific grammatical category. This approach can be used for other grammatical categories, in Albanian or in any other language.
Except this contribution, it shows that a machine learning algorithm can be applied to completely define the two-level morphology for this language.

8. References

Agalliu, F.; Angoni, E.; Demiraj, S.; Dhrimo, A.; Hysa, E.; Lafe, E. & Likaj, E. (2002). *Gramatika e gjuhës shqipe 1*, ISBN: 99927-761-6-1, BASH, Tirana

Antworth, E. L. (1990). PC-KIMMO: A Two-level Processor for Morphological Analysis, *Summer Institute of Linguistics*, Dallas

Antworth, E. L. (1995). Introduction to PC-KIMMO, *North Texas Natural Language Processing Workshop*, University of Texas

Beesly, K. R. & Karttunen, L. (2003). *Finite state morphology*, CSLI Publications, Leland Stanford Junior University, USA

Johnson, M. (1984). A discovery procedure for certain phonological rules, *Proceedings of COLING-84*, pp. 344 - 347, ISBN: 9991746080, Stanford, CA, July 1984, ACL

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing*, Prentice Hall, ISBN: 0-13-095069-6, Pearson Higher Education, New Jersey

Kadriu, A. & Zdravkova, K. (2008). Semi-Automatic Learning Of Two-Level Phonological Rules For Agentive Nouns, *Proceedings of EUROSIM/UKSim 2008,* pp. 307-312, ISBN: 9781424431922, Cambridge, England, April 2008, IEEE

Koskenniemi, K. (1983). Two-level Morphology: A General Computational Model for Word-Form Recognition and Production, *PhD Dissertation*, Department of General Linguistics, University of Helsinki

Maleki, J. & Ahrenberg, L. (2008). Converting Romanized Persian to the Arabic Writing System, *Proceedings of LREC 2008,* pp. 2904-2908, ISBN: 2-9517408-4-0, Marrakesh, Morocco, May 2008, ELRA

Mitkov, R. (2003). *The Oxford handbook of computational Linguistics*, Oxford University Press Inc., ISBN: 978-0-19-823882-9, New York

Oflazer, K. & Nirenburg, S. (1999). Practical Bootstrapping of Morphological Analyzers, *Proceedings of CoNLL-99, Workshop at EACL'99*, pp. 143-146, Bergen, Norway, June 1999, Springer Verlag

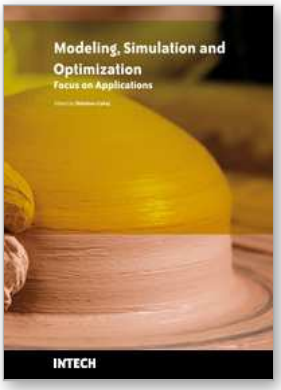
Theron, P. & Cloete, I. (1997). Automatic Acquisition of Two-Level Morphological Rules, *Proceedings of ANLP*, pp. 103 - 110, Washington, USA, April 1997, ACL

Touretzky, D. S.; Elvgren, G. & Wheeler, D. W. (1990). Phonological rule induction: An architectural solution, *Proceedings of COGSCI-90,* pp. 348–355, Hillsdale, New Jersey, August 1990, Lawrence Erlbaum Associates

Trommer, J. (1997). Eine Theorie der albanischenVerbflexion in mo_lex, *M.A. thesis*, University of Osnabruck

Trommer, J. & D. Kallulli (2004). A Morphological Tagger for Standard Albanian, *Proceedings of the LREC 2004,* pp. 201-225, ISBN: 2-9517408-1-6, Lisbon, Portugal, May 2004, ELRA

Uibo, H. (2003). Experimental Two-Level Morphology of Estonian, *8th Estonian Winter School in Computer Science (EWSCS)*, Palmse



Modeling Simulation and Optimization - Focus on Applications

Edited by Shkelzen Cakaj

ISBN 978-953-307-055-1
Hard cover, 312 pages
Publisher InTech
Published online 01, March, 2010
Published in print edition March, 2010

The book presents a collection of chapters dealing with a wide selection of topics concerning different applications of modeling. It includes modeling, simulation and optimization applications in the areas of medical care systems, genetics, business, ethics and linguistics, applying very sophisticated methods. Algorithms, 3-D modeling, virtual reality, multi objective optimization, finite element methods, multi agent model simulation, system dynamics simulation, hierarchical Petri Net model and two level formalism modeling are tools and methods employed in these papers.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Arbana Kadriu (2010). Modeling a Two-Level Formalism for Inflection of Nouns and Verbs in Albanian, Modeling Simulation and Optimization - Focus on Applications, Shkelzen Cakaj (Ed.), ISBN: 978-953-307-055-1, InTech, Available from: http://www.intechopen.com/books/modeling-simulation-and-optimization-focus-on-applications/modeling-a-two-level-formalism-for-inflection-of-nouns-and-verbs-in-albanian

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821