Abstract: - One of the most explored fields of NLP is morphology. It is important because language is productive: in any given text we will encounter text words and word forms that we haven’t seen before and that are not in a precompiled dictionary. The core task of computational morphology is to take a word as input and produce a morphonological analysis for it. There are a lot of approaches in formalizing and automatically finding rules that implement the morphology of a language. One approach is through decision lists. Decision lists can be employed as a representation language for a wide range of tasks. Clog is a system that is based on this logic and that has been primarily developed with natural language applications in mind. So far, Clog has been successfully employed for morphology learning tasks. The research presented in this paper is about machine learning of the rules of nouns morphology in Albanian. Starting point of the research presented here are 7 lists, each consisting of 6000 pairs of feminine nouns in Albanian. The pairs in the lists hold the noun’s base form and inflected form. As base word is considered the definite singular of nominative.

Key-Words: - morphological segmentation, nouns, inductive logic programming, Albanian language

1 Introduction
One of the most explored fields of NLP is morphology. It is important because language is productive: in any given text we will encounter text words and word forms that we haven’t seen before and that are not in a precompiled dictionary. The core task of computational morphology is to take a word as input and produce a morphonological analysis for it. Morphotactics defines concatenation of the principal morphemes of a word, and it is typically described through finite state automata. 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 [1]. Unsupervised learning refers to the computational task of making inferences (or acquiring knowledge) about the structure that lies behind some set of data without any direct access to that structure. In the case of unsupervised learning of morphology, and the possibilities of morpheme-combinations, for a set of words, based on no knowledge whatsoever of the language from which the words are drawn. Linguistica is a program which can be used to explore the unsupervised learning of natural language, with primary focus on morphology, which is to say, word-structure [2]. The issues from computational morphology are treated using the concepts of inductive logic programming also. One of the first attempts was done by Mooney&Califf. They focus on the problem of synthesis of the past tense of English verbs (one of the first examples of learning morphology with ILP) using the ILP system FOIDL [3]. Another open-source toolkit for unsupervised learning is Clog[1]. It is a simple but efficient first-order decision list learning system. Decision lists can be employed as a representation language for a wide range of tasks. Clog has been primarily developed with natural language applications in mind. So far, Clog has been successfully employed for morphology learning tasks. Thus, we can expect that the model is in fact universal and it may be possible to describe Albanian nouns morphology in this framework as well.

2 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 1 shows the declension of the masculine noun lis (type of tree). Table 2 shows the declension of the feminine noun fushë (field).

The base form is considered the indefinite article nominative [5]. 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).

Table 1. The declension of the masculine noun lis (tree)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom.</td>
<td>lis (tree)</td>
<td>lisa (trees)</td>
<td>lisi (the tree)</td>
<td>lisat (the trees)</td>
</tr>
<tr>
<td>Gen.</td>
<td>lisi</td>
<td>lisave lisisit lisave</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dat.</td>
<td>lisi</td>
<td>lisave lisisit lisave</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.</td>
<td>lisi</td>
<td>lisa lisisin lisat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abl.</td>
<td>lisi</td>
<td>lisash lisisit lisave</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The declension of the feminine noun fushë (field)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom.</td>
<td>fushë (field)</td>
<td>fusha (fields)</td>
<td>fusha (the field)</td>
<td>fushat (the fields)</td>
</tr>
<tr>
<td>Gen.</td>
<td>fushë</td>
<td>fushave fushës fushave</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dat.</td>
<td>fushë</td>
<td>fushave fushës fushave</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.</td>
<td>fushë</td>
<td>fusha fushën fushat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abl.</td>
<td>fushë</td>
<td>fushave fushës fushave</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 Problem Solution

CLOG is a simple but efficient first-order decision list learning system. First-order decision lists are essentially Prolog clauses each ending with a cut (!). Decision lists can be employed as a representation language for a wide range of tasks. CLOG has been primarily developed with natural language applications in mind. So far, CLOG has been successfully employed for morphology learning tasks. The training examples should be encoded as Prolog facts in a Prolog readable file. You need to create an application specific problem specification file that will permit CLOG to understand the training examples and correctly generate decision lists from it [4].

The ILP (inductive logic programming) wording of the problem is as follows: a logic program has to be learned defining the relation plural (Singular_Noun, Plural_Noun), where Singular_Noun is an orthographic representation of the singular of a noun and Plural_Noun is an orthographic representation of plural. Singular_Noun is the input and Plural_Noun the output argument. Given are examples of input/output pairs, such as plural([y,e,a,r],[y,e,a,r,s]) and plural([m,a,n],[m,e,n]). The program for the relation plural uses the predicate split(X,Y,Z) as background knowledge: this predicate splits a list of letters X into two lists Y and Z, e.g., split([y,e,a,r,s],[y,e,a,r],[s]).

2.1 Dataset

Starting point of the research presented here are 7 lists, each consisting of 6000 pairs of feminine nouns in Albanian. The pairs in the lists hold the noun’s base form and inflected form. As base word is considered the definite singular of nominative. As we explained in the second part of this paper, the same form can be used for several cases. This is so because some cases are defined using prepositions, and other cases are defined from the context. Since there are 8 cases with different form, we created 7 lists of pairs (the first case is used as a base form).

2.2 Implementation and Results

First step in the implementation was the preparing of the training examples encoded as Prolog facts in a Prolog readable file, where each word is described as a list of characters. For the base noun atmosfera (the atmosphere) the facts are as follows:

```prolog
case([a,t,m,o,s,f,e,r,a],[a,t,m,o,s,f,e,r,e]).
case([a,t,m,o,s,f,e,r,a],[a,t,m,o,s,f,e,r,ë,s]).
case([a,t,m,o,s,f,e,r,a],[a,t,m,o,s,f,e,r,ë,n]).
case([a,t,m,o,s,f,e,r,a],[a,t,m,o,s,f,e,r,a]).
```
The next step running the system for each list of the training examples. The training lists consist of 200 pairs each.

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 [5].

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

1) dite (S_indef_gen&dat&rrj)
2) ditës (S_def_gen&dat&rrj)
3) ditën (S_def_acc)
4) ditë (S_indef_nom&acc)
5) ditët (P_def_nom&acc)
6) ditëve (P_indef-gen&d_def_gen&dat&abl)
7) ditësh (P_indef_abl)
8) dita (S_def_nom)

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 gained results are as follows:

Table 3. The accuracy of the gained rules for 7 different case forms

<table>
<thead>
<tr>
<th>Case</th>
<th>Number of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

The gained rules are printed to an output file. For the first case, these rules were gained:

```prolog
case([a, t, m, o, s, f, e, r, a], [a, t, m, o, s, f, e, r, a]).
case([a, t, m, o, s, f, e, r, a, v, e]), !.
case([a, t, m, o, s, f, e, r, a, s, h]).
```

The file that contains the rules and exceptions is then used as an input to the testing system. Testing is done over the remaining examples, using a Perl script that uses AI:Prolog to query every fact in a file:

```perl
while($fact = <INPUT>) {
    chomp($fact);
    my $query = Term->new($fact);
    my $engine = Engine->new($query, $database);
    Engine->formatted(1);
    print OUTPUT $engine->results,"\n";}
```

The gained average accuracy is 97%.

4 Conclusion
The ILP model has been proved successful for formalizing the morphology of very different languages. This research presents the first logic description for Albanian. We successfully applied the ILP system CLOG in learning rules for morphological segmentation of Albanian nouns. The gained high accuracy for this category furthers in using this approach to learn rules for other grammatical categories as well.

References: