A Morphological Tagger for Standard Albanian

Jochen Trommer
Institute of Cognitive Science
Katharinenstrasse 24
D-49074 Osnabrück
jtrommer@uos.de

Dalina Kallulli
Telecommunications Research Center
Donau-City-Strasse 1
A-1220 Vienna
Kallulli@ftw.at

Abstract

In this paper, we present a morphological tagger for standard Albanian intended as a component of an annotation tool in the context of the Albanian Corpus Initiative. The analyzer uses off-line components for generating sub-regular and irregular word forms based on the verb inflector described in Trommer (1997) and simple morphological rules for main inflectional patterns. Part of the tagger are a tokenizer, a complete tagset for Albanian and full form lexica for pronouns and irregular open-class elements.

Keywords: Morphological analysis, part-of-speech tagging, Albanian

1 Introduction

Due to the political situation, there has been few research on Albanian in contemporary linguistic frameworks and virtually no work in corpus linguistics. In this paper, we present a morphological tagger which is intended as a main component of a complete part-of-speech tagger to become part of an large annotated text corpus for standard Albanian. Under a theoretical point of view, tagging Albanian is especially challenging since it has extremely rich inflectional paradigms. Thus, a verb might have up to 100 different forms. A further complication are different inflectional patterns for lexemes of the same syntactic category: Verbs fall in 53 different conjugational (Buchholz et al., 1992), while the assignment of plural affixes to noun stems does not follow from any known systematic principle.

We assume that a morphological tagger assigns to all word tokens in a text a set of morphological tags which encode the morphological features of specific word forms such as part of speech, case tense, etc. In a full-fletched part-of-speech tagger, this is supposed to be complemented by a morphological disambiguator which chooses from each such tagset a unique tag for each token given its context (figure 1).

Here is an overview of the rest of the paper: In

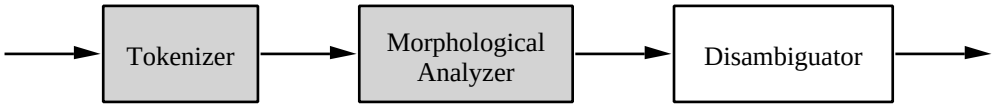

Figure 1: Architecture of PoS Tagger (shaded components are implemented in the system)

section 2, we give a short survey of Albanian inflection. Section 3 introduces the tokenizer, and section 4 describes the tagset we use in our system. The morphological analyzer is explained in section 5 and the architecture of the lexicon in section 6. Section 7 contains some remarks on the implementation of the tagger, and in section 8, we present preliminary results on the accuracy of the analyzer. Finally, in section 9, we discuss further prospects of the system.

2 Albanian Inflection

We discuss here only the inflection of open-class elements which are implemented by rules in our system. Pronominal elements show also interesting inflectional patterns[1], but these are captured by listing in a full-form lexicon.

2.1 Adjectives

Apart from few irregular lexemes, adjectives fall into five different inflectional classes which use the affixes *-e* (feminine gender), *-a* (feminine plural), *-ë* (masculine plural) or zero marking in different partially overlapping distributions. As shown in Trommer (2001), this complex allomorphy pattern can be derived by rules from the phonological shape and the morphological constituency of adjectival stems.

[1]See e.g. Trommer (2000) on the so-called preposed article and possessive pronouns.

2.2 Nouns

Nouns are inflected for number (singular, plural), case (nominative, dative, accusative, ablative)[2] such as in *shtëpi-a-ve-t*, houses-PL-ABL-DEF, 'from the houses'. While definiteness and case marking is quite regular, i.e. predictable on the basis of phonology, stem gender and number, the choice of the plural suffix (*-ë*, *-Ø*, *-e*, or *-a*) is largely unpredictable.

2.3 Verbs

Verbs are the most complex area of Albanian inflection. Apart from three different tenses (present tense ,aorist, imperfect)[3] and two different voices (active and non-active), there are five different moods (indicative, subjunctive, optative, imperative and admirative). Allomorphy in verbal inflection is partly phonologically governed. Thus verbs ending in vowels form the 1st person aorist with *-va* (e.g. *puno-va*, 'I worked') while stems ending in consonants take *-a* (e.g. *hap-a*, 'I opened'). More complex is the division of verbs in different inflectional classes which results partly in different allomorphs of

[2]Traditional Albanian grammars also assume a genitive case which however falls together in all forms with the dative.

[3]In addition to these synthetic tenses, there are two analytic tenses: future (formed with the present subjunctive and the particle *do* and the perfect formed with the participle form and finite forms of the auxiliaries *kam*, 'have', and *jam*, 'be'.

affixes (e.g. for 1sg *-j* in *mëso-j*, 'I learn' and *-m* in *the-m*, 'I say'), partly in modification of the final vowels and/or consonants of the verb stems (e.g. *vret*, 'he kills' vs. *vris-ni*, 'you (pl.) kill'). A detailed analysis of Albanian verb inflection can be found in Trommer (1997)

3 The Tokenizer

The tokenizer is a small Python script which crucially isolates word forms, punctuation marks and numbers, etc. Note that we treat some punctuation marks, such as "." (dot) and "'" (apostrophe) as a single token in some circumstances and as part of a more complex token in others. Thus *s'punon*, ('(s)he doesn't work') results in three tokens "*s*" ('not'), "'" and *punon* ('s(he) works'), while the clitic group *t'i* ('to you them') is analyzed as one token, since we store clitic groups showing many idiosyncrasies as full forms in the lexicon.

4 The Tagset

Since to our knowledge there is no published tagset for Albanian, we had to develop a complete tagset for the language.[4] As in the EAGLE guidelines standard (Leech and Wilson, 1999), tags consist of sets of attribute-value pairs. However, attributes and values are designed to fit optimally the description of Albanian and to allow a perspicuous abbreviatory notation (see below). (1a) shows a representative tag for a feminine definite (i.e., bearing an article suffix) singular common noun. To enhance legibility, we use for most practical purposes the abbreviatory notation exemplified in (1b), where all binary-valued attribute-value pairs are written by prefixing "+" or "-" to the attribute (e.g. "+def" instead of "def:+") and attributes are omitted for all other pairs (e.g. "n" instead of "cat:n"). This is possible since each (non-binary) value in our tag set corresponds to a single attribute.

(1) **Short Notation for Tags**

a. [cat:n case:nom num:sg def:+ gen:fem]
b. [n nom sg +def fem]

In addition to standard part-of-speech categories, we use "pa" for preposed articles, grammatical morphemes unique to Albanian occurring with most adjectives and possessor phrases and "ptl" for a specific class of verb-adjacent particles (e.g. future *do*).
The implementation uses intermediate representations to collapse different tags for syncretic forms of the same lexeme. Thus, the indefinite nominative and singular of all nouns is identical to the corresponding accusative form. Instead of writing the two tags (2a,b) we use the tag (2c):

(2) **Collapsed Tags**

a. [n **nom** sg -def fem]
b. [n **acc** sg -def fem]
c. [n {**nom,acc**} sg -def fem]

5 The Analyzer

The morphological analyzer consists of three components, an operative lexicon stored in a database, a set of morphological rules and a rule

[4] See http://sol.cl-ki.uni-osnabrueck.de/~atag/ for a complete list of the tagset.

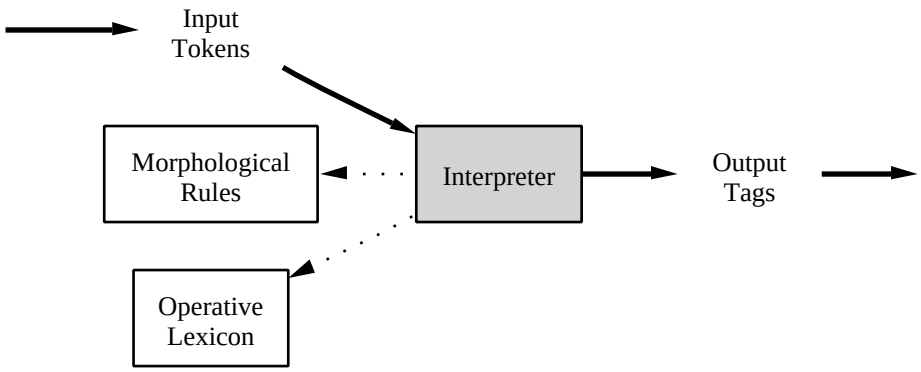

Figure 2: Structure of the morphological analyzer

interpreter (figure 5). The operative lexicon itself is partially precompiled by rules, but this happens off-line (see section 6 for discussion). Here, we will focus on the format of morphological rules and their application.

5.1 Morphological Rules

Following a long tradition in descriptive grammar and generative rule-based approaches to morphology (e.g Anderson, 1992), the morphological rules we use denote relations between input (lexicon) and output (derived) forms, where forms are ordered pairs of strings (e.g. "punoj') and tags (e.g. "[v]"). (3) shows as an example the lexeme *punoj*, 'work' and its 2nd/3rd person singular form *punon*:

(3) **Input-Output Pair**

Input:	<punoj, [v] >
Output:	<punon, [v {2 3} sg ind pres] >

Rules are quintuples of the form $<left\_context$, $remove$, add, $lexicon\_category$, $tag>$, where $left\_context$ and $remove$ are regular expressions and all other components strings. $lexicon\_category$ specifies the category tag of the entry in the operative lexicon and tag the resulting tag. add is the suffix which is added to the stem after removing an expression corresponding to (stem-final) $remove$ to get the word form. The rule can only be applied if the suffix of the input stem corresponding to $remove$ is preceded by a string matched by $left\_context$.

Figure 3 contains a slightly simplified example of a morphological rule. This rule deletes a final *j* ($remove$) from an item which has the lexicon category "[v]" if *j* is preceded by a vowel ($left\_context$), and adds *n* instead which gets the tag "[v 2 3 sg ind pres]". Figure 4 shows how the rule applies to the example pair from (3).

The morphological rules we use do not differentiate between phonology and morphology. Thus the fact that the 1st person singular aorist suffix for verb stems ending in a consonant is *-a* (e.g. *hap-a* 'I opened', while it is *-va* after vowels (e.g *pi-va*, 'I drank' is not captured by

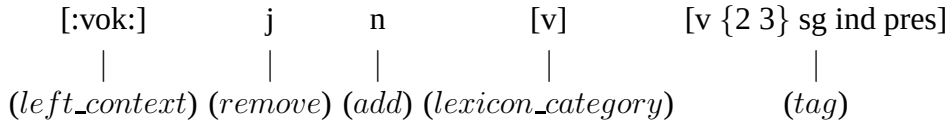

Figure 3: Example for a morphological rule

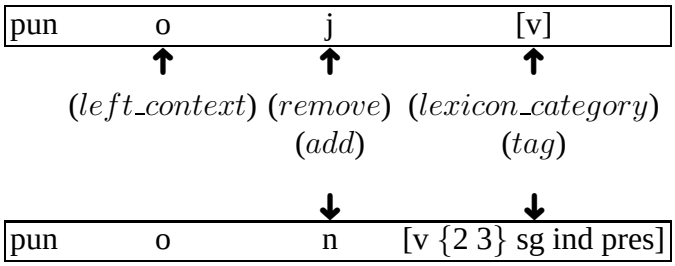

Figure 4: The rule from figure 3 applied to "punon [v]"

a separate phonological rule, but simply by two different morphological rules:

(4) **Morphological Rules for 1sg aorist**

a.	[:vok:]	0	va	[v]	[v 1 sg aor]
b.	[:kons:]	0	a	[v]	[v 1 sg aor]

In approaches to morphological analysis such as two-level morphology (see Karttunen and Beesley, 2001, and references cited there) which separate phonology and morphology, an alternative to assuming two different affixal items for 1sg aorist would be to assume just one (say *-va*) and derive the other form by a phonological rule (here: delete *v* after a consonant). We think that these approaches are well-motivated in languages with rich sandhi phenomena such as Finnish, but lead to unnecessary complexity in a language like Albanian which shows – at least at the orthographic level – few such processes.

5.2 The Rule Interpreter

Recall that morphological rules, although we have discussed them as devices to derive word forms, are declarative statements on relations between lexicon entries and word forms. In fact, our rule interpreter uses these rules to infer possible lexical entries for a given word form. It transforms the $left\_context$ and add parts of each rule into one regular expression. For each word form which matches this expression for a rule R with suffixes S, it combines the remaining prefixes P of the word form with the $remove$ parts compatible by R with S to get a set of potential lexicon forms which are then checked against the lexicon data base. Since

there is usually at most one analysis a rule assigns to a word form and few rules matching a given suffix of a word form, search space is small.

6 Lexicon Construction

Morphological analysis in our system is especially simple since each corresponding pair of lexical entries and word forms is related by exactly one rule. In other words, there is no iterative rule application. This is possible since the operative lexicon which serves as the basis for rule application is itself constructed by rules from different source lexica to derive e.g. singular and plural stems for nouns.

There are three source lexica for the operative lexicon: 1) the *full-form lexicon* 2) the *stem lexicon* and 3) the *regular lexicon*. The *regular lexicon* contains stems formed by redundancy rules from the *base lexicon* (see subsection 6.2), the *stem lexicon* irregular stems which however form the basis of additional morphological rules (e.g. for the irregular noun plural *duar*, 'hands' to which still case and definiteness affixes can be attached) and the *full-form lexicon* complete word forms with tags which are accessed by a default morphological rule also responsible for treating uninflected lexicon entries.

Since entries for a given lexeme are all together in one of these lexica, there is a simple algorithm to construct the operating lexicon from the three source lexica (5). $A \approx A'$ denotes here the relation of two lexicon entries which refer to the same lexeme.

(5) **Lexicon Formation Algorithm**

for all lexemes l in reg_lex:
  if $\exists$ entries $l_1 \ldots l_n \approx l$ in full_form_lex**:**
    add $l_1 \ldots l_n$ to operating_lex
  else if $\exists$ entries $l_1 \ldots l_n \approx l$ in stem_lex**:**
    add $l_1 \ldots l_n$ to operating_lex
  else:
    add l to operating_lex

6.1 Exception Lexica

While the exception lexica (i.e., the stem lexicon and the full-form lexicon) are for the most part static lists of stems and full forms, irregular verb forms in the full-form lexicon are created by the generation tool for Albanian verb forms described in Trommer (1997) based on **mo_lex**.

6.2 Redundancy Rules

Redundancy Rules apply to the items of the base lexicon which contains a list of all basic stems with part-of-speech tags to derive the full list of regularly formed stems in the regular lexicon on the basis of phonological and morphological properties of the base stems. For example Albanian nouns ending in *-im*, regularly take the plural affix *-e*. Thus, a redundancy rule creates for each noun stem in the base lexicon which ends in *-im* a plural stem with the suffix *-e* in the regular lexicon. Redundancy rules are directly implemented as Python scripts.

7 Implementation

The morphological tagger is implemented under SuSe Linux 8.0. using Python 2.1 and MySQL

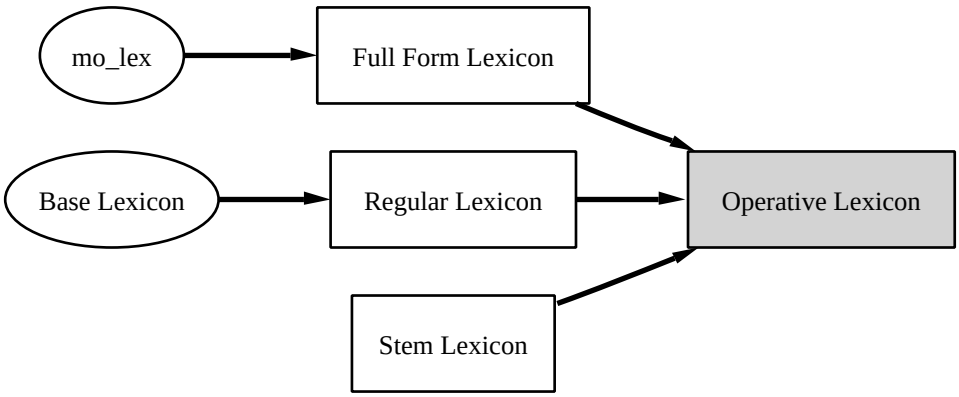

Figure 5: Lexicon Construction

11.18. There are currently 340 morphological rules. The operative lexicon contains 53054 entries. The base lexicon for open-class element is mainly based on the Albanian word list from the ECI/MCI multilingual corpus CD[5]. There is a web interface to the morphological analyzer under http://sol.cl-ki.uni-osnabrueck.de/~atag/.

8 Evaluation

Work on the tagger is still in progress. The rules by now follow mainly the descriptions in Buchholz and Fiedler (1987) and Buchholz et al. (1992) which give the most detailed description of Albanian morphology. It remains necessary to optimize the analyzer with respect to running text from corpora. To test the accuracy of the tagger in its current state, we tagged two texts representing different text sorts containing each 500 word tokens (an initial part of a novel (Kadaré, 1990) from the ECI/MCI multilingual corpus CD and part of a news article from Albanews[6]) by hand and compared the results to the tags produced by the tagger. To quantify accuracy, we use the standard measures *precision* and *recall*, where "precision is the number of correct token-tag pairs that is produced, divided by the total number of token-tag pairs that is produced, and recall is the number of correct token-tag pairs that is produced, divided by the number of correct token-tag pairs that is possible." (van Halteren, 1999:82) The table in (6) shows the results for the tokens in the two texts (Text1 = Albanews, Text2 = Kadaré, Both = both texts concatenated). "all" stands for the complete texts including punctuation marks, "words" for the texts with punctuation marks removed. (7) shows the corresponding measures for word types.

(6) **Accuracy for Tokens**

		precision	**recall**
Text1	all	98% (890)	95% (919)
Text2	all	97% (896)	95% (920)
Both	all	97% (1786)	95% (1839)
Text1	words	98% (833)	94% (861)
Text2	words	97% (791)	94% (815)
Both	words	97% (1624)	94% (1676)

[5] http://www.elsnet.org/resources/eciCorpus.html

[6] http://listserv.acsu.buffalo.edu/archives/albanews.html, message 61 of week1, November 1997.

(7) **Accuracy for Types**

		precision	**recall**
Text1	all	96% (389)	92% (409)
Text2	all	97% (425)	93% (444)
Both	all	97% (719)	92% (758)
Text1	words	96% (385)	92% (404)
Text2	words	97% (419)	92% (438)
Both	words	97% (713)	92% (751)

While we have not done a detailed error analysis so far, a first survey suggests that errors, especially in recall are mainly due to missing lexicon entries in the system, most of them names, but also nouns and verbs.

9 Further Prospects

We expect further improvement of the taggers accuracy from a substantial revision of the lexicon for open-class lexemes. The next step we plan is the development of a statistical disambiguator to get a full-fletched part-of-speech tagger, which is intended as a contribution to a large-scale annotated text corpus for Albanian.

References

Anderson, S. R. (1992). *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Buchholz, O. and Fiedler, W. (1987). *Albanische Grammatik*. Leipzig: VEB Verlag Enzyklopädie.

Buchholz, O., Fiedler, W., and Uhlisch, G. (1992). *Wörterbuch Albanisch-Deutsch*. Langenscheidt Verlag Enzyklopädie: Leipzig.

Kadaré, I. (1990). *Koncert në fund të dimrit*. Shtëpia Botuese ”Naim Frashëri: Tiranë.

Karttunen, L. and Beesley, K. R. (2001). A short history of two-level morphology. Xerox Palo Alto Research Center and Xerox Research Centre Europe.

Leech, G. and Wilson, A. (1999). Standards for tagsets. In van Halteren, H., editor, *Syntactic Wordclass Tagging*, chapter 5, pages 55–80. Kluwer Academic Publishers.

Trommer, J. (1997). Eine Theorie der albanischen Verbflexion in **mo_lex**. M.A. thesis, University of Osnabrück.

Trommer, J. (2000). The post-syntactic morphology of the Albanian pre-posed article. In *Proceedings of the third conference on South-Slavic and Balkan languages, Plovdiv, September ’99*. Evidence for Distributed Morphology.

Trommer, J. (2001). Phonologically conditioned allomorphy in Albanian adjectives. Ms., University of Osnabrück.

van Halteren, H. (1999). Performance of taggers. In van Halteren, H., editor, *Syntactic Wordclass Tagging*, chapter 4, pages 81–94. Kluwer Academic Publishers.