A Morphological Analyzer for Standard Albanian

Jochen Trommer*, Dalina Kallulli*

*Institute of Cognitive Science
Katharinenstrasse 24, D-49074 Osnabrück, Germany
jtrommer@uos.de

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

Abstract
In this paper, we present a morphological analyzer 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 analyzer are a complete tagset for Albanian and full form lexica for pronouns and irregular open-class elements.

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 analyzer 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 classes (Buchholz et al., 1992), while assignment of plural affixes to noun stems does not follow from any known principle.

We assume that a morphological analyzer 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.

Here is an overview of the rest of the paper: In section 2, we give a short survey of Albanian inflection. Section 3, describes the tagset we use in our system. The morphological analyzer is explained in section 4, and the architecture of the lexicon in section 5. Section 6, contains some remarks on the implementation, and in section 7, we present preliminary results on the accuracy of the analyzer.

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, 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), -d (masculine plural) or zero marking in different partially overlapping distributions. As shown in Trommer (2001), this complex allomorphy can be derived by rules from the phonological shape and morphological constituency of adjectival stems.

2.2. Nouns
Nouns are inflected for number (singular, plural), case (nominative, dative, accusative, ablative)† and definiteness 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 (-d, -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)‡ 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 -a (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 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 Tagset
Since to our knowledge there is no published tagset for Albanian, we had to develop a complete tagset for the lan-

---

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

‡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’.
guage. 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 (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]

4. The Analyzer

The morphological analyzer consists of three components, an operative lexicon stored in a database, a set of morphological rules and a rule interpreter (figure 4.). The operative lexicon itself is partially precompiled by rules, but this happens off-line (see section 5. for discussion). Here, we will focus on the format of morphological rules and their application.

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

(4) Morphological Rules for 1sg aorist

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

4.2. 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 database. 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.

5. 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 ap-
plication 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 reg-
ular lexicon. The regular lexicon contains stems formed
by redundancy rules from the base lexicon (see subsection
5.2.), the stem lexicon irregular stems which however form
the basis of additional morphological rules (e.g. for the ir-
regular noun plural duar, ‘hands’ to which still case and
definiteness affixes can be attached) and the full-form lexi-
con 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

5.1. 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.

5.2. 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 morpho-
logical 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 imple-
mented as Python scripts.

6. Implementation

The morphological analyzer is implemented under SuSe
Linux 8.0. using Python 2.1 and MySQL 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
4. There is a web interface to the morphological analyzer under
http://sol.cl-ki.uni-osnabrueck.de/~atag/.

7. Evaluation

Work on the analyzer 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 de-
tailed description of Albanian morphology. It remains nec-
essary to optimize the analyzer with respect to running text
from corpora. To test the accuracy of the analyzer in its cur-
rent 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 cor-
pus CD and part of a news article from Albanews5) by hand
and compared the results to the tags produced by the ana-
lyzer. To quantify accuracy, we use the standard measures

4http://www.elsnet.org/resources/eciCorpus.html
5http://listserv.acsu.buffalo.edu/archives/albanews.html, mes-
  sage 61 of week1, November 1997.
**Figure 3:** The rule from figure 3 applied to “pun [v]”

**Figure 4:** Lexicon Construction

*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) |

(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.

8. References

Anderson, Steven R., 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Buchholz, Oda and Wilfried Fiedler, 1987. *Albanische Grammatik*. Leipzig: VEB Verlag Enzyklopädie.

Buchholz, Oda, Wilfried Fiedler, and Gerda Uhlisch, 1992. *Wörterbuch Albanisch-Deutsch*. Langenscheidt Verlag Enzyklopädie: Leipzig.

Kadaré, Ismail, 1990. *Koncert në fund të dimrit*. Shëtqia Botuese Naim Frashëri: Tiranë.

Karttunen, Lauri and Kenneth R. Beesley, 2001. A short history of two-level morphology. Xerox Palo Alto Research Center and Xerox Research Centre Europe.

Leech, Geoffrey and Andrew Wilson, 1999. Standards for tagsets. In Hans van Halteren (ed.), *Syntactic Wordclass Tagging*, chapter 5. Kluwer Academic Publishers, pages 55–80.

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

Trommer, Jochen, 2001. Phonologically conditioned allomorphy in Albanian adjectives. Ms., University of Osnabrück.

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