Electronic Dictionaries and Transducers for Automatic Processing of the Albanian Language

Odile Piton [1], Klara Lagji [2], Remzi Përnaska [3]

[1]University Paris1 Panthéon-Sorbonne, [3]University of Poitiers France
[2]University of Tirana Albania

Abstract. We intend on developing electronic dictionaries and Finite State Transducers for the automatic processing of the Albanian Language. We describe some peculiarities of this language and we explain how FST and generally speaking NooJ's graphs enable to treat them. We point out agglutinated words, mixed words or 'XY' words that are not (or cannot be) listed into dictionaries and we use FST for their dynamic treatment. We take into consideration the problem of unknown words in a lately reformed language and the evolving of features in the dictionaries.

Keywords: Morphological Analysis, Electronic Dictionary, Finite State Transducer, Albanian Language, Automatic Processing of Open Lists of Words.

1 Introduction

The evolution of powerful of computers and the increased size of central and external memory storage make it possible to process natural language with dictionaries. However, "Large-scale lexical acquisition is a major problem since the beginning of MT" [1]. Boitet suggests mutualizing creation and usage of lexical resources. The automatic processing of Albanian is not yet developed [7] [9] [10]. So this work focuses on syntactic and basic semantic features. The completion of semantic information requires many steps and lots of human labour.

A lot of papers have been written on dictionaries for NLP. In this one, we won't describe the usual work of generating the list of known words with their category and features, building tools to describe and generate flexed words. Instead, we address some advanced problems of this language and we explain how Finite State Transducers -FST- described with NooJ's graphs enable to solve them[1]. Nooj furnishes tools for automatically constructing words by transducers: there are inflectional grammars, morphological grammars and syntactic grammars.[2].

Dictionaries are two levelled structures [14]. "The words of the natural language are complex entities and they sometimes have a much more elaborated internal structure… they often join other lexical elements to form wider units as compound words or expressions [15]." Language is a creative process. The dictionaries needed by automatic treatment must register basic vocabulary and be associated with tools, with creative paradigms, that compute the new words of the constructed part [2]. The choice of the set of syntactic and semantic features is important.

1 See http://www.nooj4nlp.net/ for information on NooJ.
2 See http://marin-mersenne.univ-paris1.fr/site_equipe_mm/O_Piton/Piton.html for more information.

2 Some Properties of Albanian

The late reform, in 1972, has lead to standard Albanian Literary Language. A linguistic move took place in order to "unify" the two Albanian dialects: Gheg and Toskë. If the official language seems to be the language of the media [6] and the language of the schools, it is not the language of every Albanian speaker. This language is still subject to variation. Two usual assumptions are not true for Albanian: *all the stems are not in dictionaries and syntactic information is often lacking*. This gives an unusual importance to the treatment of "unknown words". We must note that lot of verbal forms are not plain forms but "analytic forms". We cannot build the automated processing of Albanian using a bottom-up method. We develop specific tools for analytic forms.

Our first data was a paper dictionary. As it has been noted: "Structure of entries in dictionaries varies considerably, inside one dictionary as well as between different dictionaries: it seems that any type of information can be found in any position in a dictionary. Nevertheless, in spite of these variations, human readers are able to interpret the entries easily and this, without needing to consult introductive explanations. It is clear that there are several principles and underlying regularities."[14] These regularities are made of slashes, commas, parenthesis and hyphens, which surround grammatical, linguistic and technical information. We have used the Albanian-French dictionary of the book "Parlons Albanais" [4]. We have observed its format. Part of speech: noun, verb, etc. is not always written. We have developed heuristic to infer it.

2.1 About Digraphs.

Albanian alphabet has 36 letters. It uses the Latin alphabet. 9 letters are written with "digraphs" or double characters: *dh gj ll nj rr sh th xh zh*, but they must be considered as one letter.

Some regular paradigms to make stem allomorphs have to be redefined. Some nouns construct an allomorph in loosing *ë* and receiving *a*. E.g. *motër → motra*. The rules "*go left one letter*" then delete one letter has to be redefined for *vjehërr → vjehrra* as "*go left two chars*" according to the fact that '*rr*' is one letter with two characters. The whole set of words that obey the regular paradigm, has to be dispatched between the '1 char.' vs. '2 char.' paradigms.

Some words like son *bir* and devil *djall* drop their last letter and receive *j* in the plural. But the last consonant of *djall* is a double letter, so we need two paradigms. *'Delete 1 letter* vs. *delete 2 letters* then insert j' *bir → bij* and *djall → djaj*.

2.2 About Verbal System

A verb can have one or two forms: *Z* active form and/or *Z-hem* not active forms: e.g. *laj* (to wash) and *lahem* (to wash oneself). The two forms have the same past participle and four tenses of non-active forms are build on the same tense as the active form preceded by the particle "u": e.g. *lava* (I washed-aorist) *u lava* (*I* washed myself). Some tenses use particles: *të* and *do* and some tenses include others. E.g. *laj, të laj, do të laj*, (active form) and *lahem, të lahem, do të lahem* (non-active form) are 6 different tenses. If we recognize *laj* as present tense, we don't see that it is part of *do të laj*, a form of the future tense. So we need to describe analytic forms. We have drawn corresponding graphs [11], to build one form for each person of each conjugation. Active and non active verbs are described by separate graphs (200 graphs and sub-graphs). Non active forms can be separated into subsets with syntactic and semantic features.

For some verbs, *aorist* is marked by a process known as "*ablaut*"; there is a change in vowel, e.g. '*e→o'*. We have to distinguish two subsets of verbs and to define two graphs: 'insert a, *go left 1 letter* vs. *go left2 letters*, delete 1 letter, insert o' *heq→ hoqa*, *hedh→hodha.*

2.3 About Nouns, Pronouns and Adjectives

In Albanian, nouns and pronouns are usually inflected. A great number of masculine words in the singular are feminine in the plural. So the gender is not a static property of a noun. *It is not written in the dictionary*. Declension position is at the end of the word, but for three pronouns, it is inside, e.g. *cilido, cilitdo, and cilindo*. It is the same phenomenon for '*auquel* / *auxquels'* in French: two concatenated words and the first one is flexed.

Foreign Named Entities are transcript according to Albanian phonetic: *Shekspir* Shakespeare, *Xhems Xhojs* James Joyce, *Sharl dë Gol* Charles de Gaulle. They are flexed as other nouns. So are acronyms, but they are preceded by a dash *OKB-ja*, *OKB-në*, two flexions of *OKB* (UNO).

Most adjectives, some nouns and some pronouns are preceded by a particle called article. These articles have declensions; their four forms can be *i, e, të, së*. These declensions vary according to the place of the articulated adjective or articulated noun in nominal syntagm.

Homonymy interferes with articulated words: e.g. *parë* seen; '*të parë'*, aspect, '*i parë'*, chief; and ancestor. The whole sequence must be recognized all over.

3 Dynamic Method for Albanian with NooJ

NooJ processes both simple and compound words, if their lemma is in NooJ's dictionaries. Some agglutinated words, or 'XY' built words need extra tools.

3.1 FST for Agglutinated or Mixed Forms

Agglutination and Mix for the Imperative Tense. The imperative can concatenate with the clitic complement *më*: *laj* (wash*), lani* (let you wash), *më laj* (wash me), *më lani* (let you wash me) can be used in the form *lamë* and *lamëni*. These agglutinations obey to specific rules (phonetic or category of verb) and must be described. NooJ's graphs allow us to recognize such agglutinated and mixed forms: *lamëni = lani + më.*

Graphs for Atonic (or Short) Forms. Morphological grammar allows us to process contracted words. For example, in French the word '*au'* is expanded as '*à le'*. There are similar contractions in Albanian. When there are several clitics before a verb, they are often contracted: *'të e'* into *ta* (*të* particle for subjunctive, or clitic dative and *e* clitic accusative), *'më e'* into *ma*, and '*i i'* or '*i e'* into *ia*. Clitic forms are '*më të e i na ju u*'. The contraction can be one single sequence or use an apostrophe: '*të i'→ t'i*. But some contractions are ambiguous and the FST uses extra information to disambiguate.

See a graph to expand *ma, ta* and *ia* in figure 1.

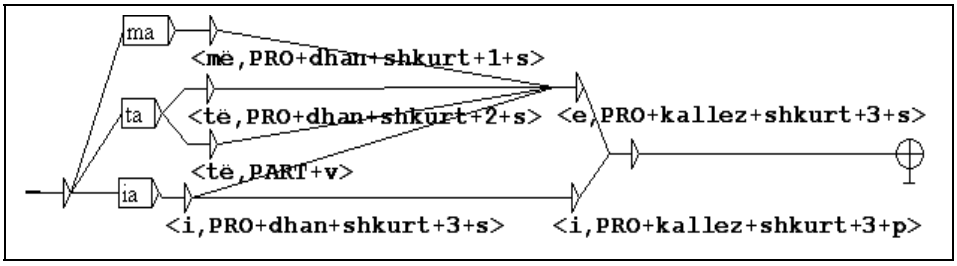

Fig. 1. Morphological graph for "decontraction" of some 'short forms' of personal pronouns: *ma*, *ia*, *ta*. Read from left to right. Last line processes 'ia' (in the box, the outputs are under the lines, they perform the translation). The graph has two ways: one way translates '*ia*' into <i> <i>, the second translates '*ia*' into <i> <e>.

3.2 FST for Open Lists

Albanian has some productive paradigms, from derivation rules to concatenation rules that are very active and produce concatenated words. It is impossible to list all XY words in dictionaries, because *the number is infinite*. This is called *open lists of words*. An interesting property is that most of the words are concatenated without contraction or modification. That makes them easy *to recognize dynamically*. See in figure 2 a graph to recognize dynamically unknown words beginning with *para*, *pa* or *nën*.

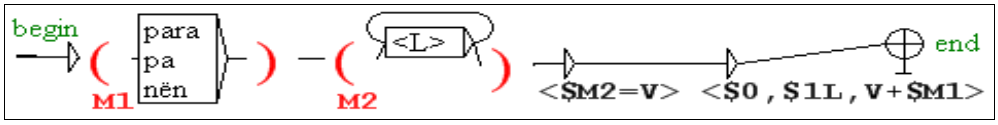

Fig. 2. Graph for XY words where X is *para*, *pa* or *nën*; Y is a verbal form. Explanation for *nënkuptoj:* the graph compares the first letters with *para*, *pa* and *nën. A variable M1* receives the result *nën*; then the loop on <L> extracts the second part *kuptoj* into M2. M2 is compared to the verbs by <$M2=V>. K*uptoj* is recognized as a verb, so the last part of the graph generates dynamically the entry: *nënkuptoj,kuptoj,V+nën*

'XY' Words Built by Concatenation. Y is a verb, noun, or adjective. X can be:

- an affix: prefix like *ç*, *sh*, zh, *pa* (negation), e.g. *krehur* (combed), *pakrehur* (not combed).
- a preposition like *nën* (under), e.g. *kuptoj* (to understand), *nënkuptoj* (to suggest).
- a noun like *flokë* (hair) e.g. *bardhë* (white), *flokëbardhë* (white haired).
- an adjective like *keq* (bad), *besim* (trust), *keqbesim* (distrust); or an adjective finished by a "o" e.g. *leksiko-gramatikore* (lexical-grammar), X is invariable, Y only takes the marks of feminine or plural: *sesioni tecniko-shkencor* (technico-scientific session) [7].

This process is recursive and very productive. With the morphologic graphs, NooJ gives a tool that allows us to recognize such constructed words on demand [12]. A morphological graph can

recognize several parts in a word and can compare them either to a specified list, or to a set of words. When the comparison is successful, the word is recognized and receives the lexical features and the syntactic features.

'XY' Words with Numbers. Concatenated cardinal numbers can be the first part of words. The second part of the 'XY' word is compared to a list or to the dictionary. E.g. number 22 *njëzet e dy* concatenated and followed by *vjeçar* (aged of) gives an adjective: *njëzetedyvjeçar* (twenty-two years old). This can also be done with words like *katesh* (floor), *mujor* (month), *orësh* (hour), etc. Words like *njëzetedyvjeçar* are not listed into dictionaries.

3.3 FST for Derivation

We noticed earlier that lot of words are not in dictionaries. Unknown words are submitted to different processes. They are compared to the derived forms (according to the stem). Their affix gives information on their POS. E.g. most of words with suffix *ik* or *ike* are adjectives, while a few are nouns. The "Inverse Dictionary of Albanian" [8] makes it possible to classify the affixes of entries. However it does not present the plural forms.

Derivational morphology is a well-known strategy to improve coverage of lexicon by affix operations by iterative process. Automatic analysis and filtration is exposed by [13]. A study of derivation is exposed by [3]. These words inherit argument structure of the base word, or part of it. The Albanian language is strongly constraint and derivation is very regular. The derivation is used for the operation of tagging unknown words. NooJ's morphological graphs construct derived words from a root and give them a category and some features.

About words with an article. We have noticed that the words that occur many times in the text should be grouped before making an entry for it. We notice a problem for articulated words that are not grouped with their article because 'unknowns' are single words.

Table 1. Example of features for the nouns.

N_Nb = s + p;	The number can be singular or plural
N_Genre = m + f + as;	The gender can be masc., fem. or neutral
N_Shquar = shquar + pashquar;	A noun can be definite or undefinite
N_Rasa = emer + rrjedh + gjin + kallez + dhan;	This is the list of names of declensions for nouns

3.4 Features in the Dictionaries

This is necessary to be able to evolve and/or improve the dictionaries. An important part of the work is to organize the syntactic and semantic features and tags for each category. Whenever it is necessary to evolve syntactical and semantic tags, the dictionary will be modified. It is important to be able to adapt it easily. For example, a new study can make it necessary to define two tags instead of one. So the set of features and tags must be carefully registered and regularly updated. NooJ registers the set of features in a file called "*properties definition*". This file can be used to display lexical information as table.

4 Conclusion

It is necessary to add entries from others Albanian dictionaries to these first electronic dictionaries. We are aware that there is a lot more work to be done.

In Albania, the Computational Linguistics is not very developed.

Since 1998, a course in Computational Linguistics is organized with the students of the Department of Albanian Linguistics at the Tirana University. It aims at the sensitization of the future linguists and language teachers about new approaches of the natural language processing. At the end of the course, the students have to do a study application of this methodology to Albanian language. But, actually, no organized collaboration exists between linguists and computer scientists.

References

1. Boitet Chr. Méthode d'Acquisition lexicale en TAO: des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes. TALN 2001, 8E Conf. Sur le TALN. Tours. (2001) 249-265.
2. Carré R. Degrémont J. F., Gross M., Pierrel J.-M., Sabah G. Langage Humain et Machine, Paris, Presses du CNRS. (1991)
3. Gdaniec C., Manandise E., McCord M. Derivational Morphology to the rescue: How It Can Help Resolve Unfound Words in MT. Proceedings, Santiago de Compostela, Spain, (2001), 129-131.
4. Gut Chr., Brunet-Gut A., Përnaska R. Parlons Albanais Paris L'Harmattan Paris (1999)
5. Habert B., Nazarenko A., Zweigenbaum P., Bouaud J. Extending an existing specialized semantic lexicon. In Antonio Rubio, Navidad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation*, pages 663-668, Granada, (1998).
6. Hasani Z. Le Déclin de la Langue Albanaise? in Shekulli, 12 novembre 2002.
7. Lagji K. Etude sur le Statut du Mot en Albanais dans le Cadre des Traitements Automatiques des Langues. In Annotation Automatique de Relations Sémantiques et Recherche d'Informations: vers de Nouveaux Accès aux Savoirs. Université Paris-Sorbonne, 27-28 octobre, Paris, (2006)
8. Murzaku A. Albanian Inverse Dictionary 32,005 words http://www.lissus.com/resources_download.htm
9. Piton O., Përnaska R. Etude de l'Albanais en Vue de Construire des Outils pour son Traitement Automatique, Journées NooJ Besançon (2005)
10. Piton O., Përnaska R. Constitution de Dictionnaires Electroniques pour l'Albanais, et Grammaire du Groupe Nominal avec Nooj, Belgrade (2006)
11. Silberztein M. Dictionnaires électroniques et analyse automatique de textes. Le système INTEX. Masson Ed. Paris Milan Barcelone Bonn (1993)
12. Silberztein M. NooJ's dictionaries. Proceedings of LTC (2005), Poznan University.
13. Tsoukermann E., Jacquemin Chr. Analyse Automatique de la Morphologie Dérivationnelle, et Filtrage de Mots Possibles, Mots possibles et mots existants, Silexicales. (28-29 avril 1997) 251-259.
14. Véronis J., & Ide N. Encodage des dictionnaires électroniques: problèmes et propositions de la TEI, in D. Piotrowsky (Ed.), Lexicographie et informatique - Autour de l'informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy (29, 30, 31 mai 1995) 239-261.
15. Wehrli E. L'analyse syntaxique des langues naturelles, in Problèmes et Méthodes, Masson Ed. Paris Milan Barcelone (1997)