Conception of a NooJ module for the Automatic Processing of the Albanian language

Odile Piton

Université Paris 1

Klara Lagji

Université de Tirana/ Sorbonne-Paris IV

Plan

Introduction

Considerations on some syntactical constructions of compounds Nominal Syntagms

Platform NooJ for Albanian language

- Dictionaries
- Grammar : Nominal syntagm, and Results
- Other morphological graphs

Conclusions

I Introduction

About Albanian (NooJ 2005, 2006, 2007),(NLDB 2007)

- reform in 1972 : standard Albanian Literary Language
- to “unify” the two Albanian dialects: Gheg and Tosque
- the official language seems to be the language of the medias and the language of the school
- this language is still subject to variation
- all the stems are not in paper dictionaries
- syntactic information is often lacking

“Kush e solli Doruntinën” = “Qui a ramené Doruntine” Ismail Kadare.

14,980 words, 8738 tokens

të	5,94%	**‘particle’ / verb particle / pronoun**
e	5,79%	**‘particle’ / pronoun/ conjunction**
i	2,21%	**‘particle’ / pronoun / sometimes possessive**
total	12,94%	

“Deklarata e përgjithshme mbi të drejtat e njeriut” =

“the Universal Declaration Of Human Rights ” 1848 word forms

të	247	13,37%	**‘particle’ / verb particle / pronoun**
dhe	96	5,19%	**conjunction + noun earth + verbal form**
e	79	4,27%	**‘particle’ / pronoun/ conjunction**
në	54	2,92%	**preposition**
i	46	2,49%	**‘particle’ / pronoun / sometimes possessive**
total	522	28,25%	

particles i, e, të, së **are named articles or pronouns, depending on the context.**

- particles used for
  - before a verb as personnal pronoun
  - compound nouns
  - compound adjectives
  - some pronouns,
  - can be part of other forms

EX: **Të gjithë njerëzit lindin të lirë dhe të barabartë në dinjitet dhe në të drejta. Ata kanë arsye dhe ndërgjegje dhe duhet të sillen ndaj njëri tjetrit me frymë vëllazërimi.**

→ **In the first e sentence, the article creates compound adjectives; in the second, it is a pronoun.**

Example

- **HYRJE**
- **Mbasi njohja e dinjitetit të lindur të të drejtave të barabarta dhe të patjetërsueshme të të gjithë anëtarëve të familjes njerëzore është themeli i lirisë, drejtësisë dhe paqes në botë;**

- **PREAMBLE**
- **Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world;**

problems with i, e, të, së

- **Example of e : it is**
  - a conjunction (and)
  - an « article » for feminine comp. adjective in 5 forms
  - an « article » for masculine comp. adjective in 3 forms

    But it depends on the position of the adjective after the noun : 1rst or 2nd

    And it is different if the adjective is before the noun (less often, but possible)
  - a particule to construct the genitive of nouns
  - a particule used for fractions
  - a clitic for accusative in the singular
  - an « article » for feminine comp. nouns
  - an « article » for masculine comp. Nouns
  - before a short list of feminine human words (wife, daughter, ..) it means his or her
- **a lot more complicated for të** → ALSO used in about half of the verbal forms : for the subjective forms, for the future forms, for the conditional forms, plus the infinitive

Platform : text, dictionaries, grammars

- **Albanian dictionaries : entries + flections**

Verbs

- **transitive 2900**
- **intransitive 700**
- **« reflexive » 2100 → 190 graphs → 258000 forms**

- **Nouns**
  - **feminine 9000**
  - **masculine 8600**
  - **neutral 60 → 400 graphs → 523000 forms**
  - **+ Topological Names 3000 → 45000 forms**

- **Adjectives**
  - **9900**
  - **> 50 % with a particle → 28 graphs → 38460 forms**
  - **+ Topological adjectives 834 → 3328 forms**

- **Pronouns**
- **Adverbs → 1541**
- **Prep → 100**

Text : the **Universal Declaration Of Human Rights**

- 91 paragraphs
- 1837 word forms
- 1800 annotations

V A N A CONJC A

është [i mundur] [zhvillimi [i lirë] dhe [i plotë]]

N (genitive) Poss.

[i [personalitetit [të tij]]]

→ **the main deal is to construct grammars**

Method for nominal syntagm

- Rather than creating a lot of analyses that have to be eliminated later
  - We adopt a bottom-up method
  - We recognise the nominal syntagm (Noun followed by adjective)
  - We consider the specific forms of the « article »
  - We recognise both simple adjectives and compound adjectives

- So we have a grammar for

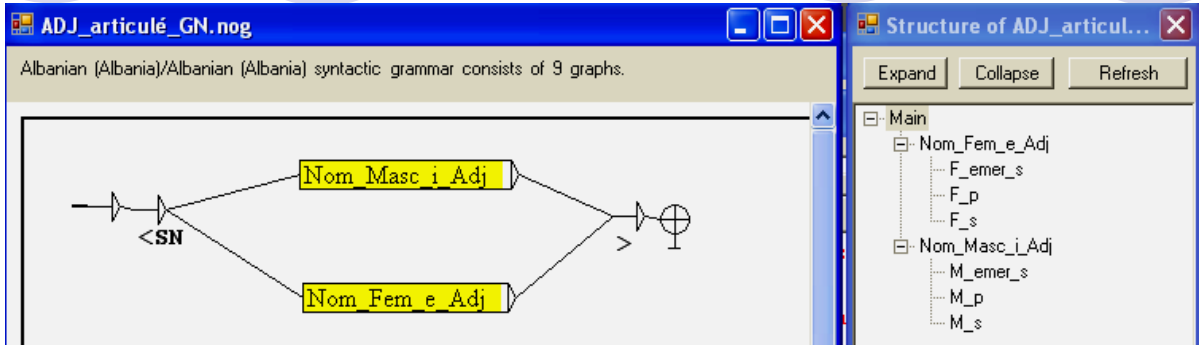
  A noun + an adjective
  A noun + a first adjective + a second adjective
  (some differences between the articles)
  A noun + a first adjective + a second adjective+ a third adjective

Nominal syntagm : first step

- **Simple adjectives are recognised by the dictionaries : 4 forms and no declension (except when they are anteposed)**
  - **Example : adjective Tunisian**
    - **tunizianë,tunizian,A+Hum+m+p**
    - **tuniziane,tunizian,A+Hum+f+s**
    - **tuniziane,tunizian,A+Hum+f+p**
    - **tunizian,A+Hum+m+s**
- **Compound adjectives receive the category AIE + feature ie**
  - **Example : adjective (i,e)barabartë: barabartë,AIE+FLX=Adj_Ba+ie**
    - **The flection creates four forms**
      - **barabartë,AIE+FLX=Adj_Ba+ie+m+s**
      - **barabartë,AIE+FLX=Adj_Ba+ie+m+p**
      - **barabartë,AIE+FLX=Adj_Ba+ie+f+s**
      - **barabarta,AIE+FLX=Adj_Ba+ie+f+p**

Nominal syntagm : second step

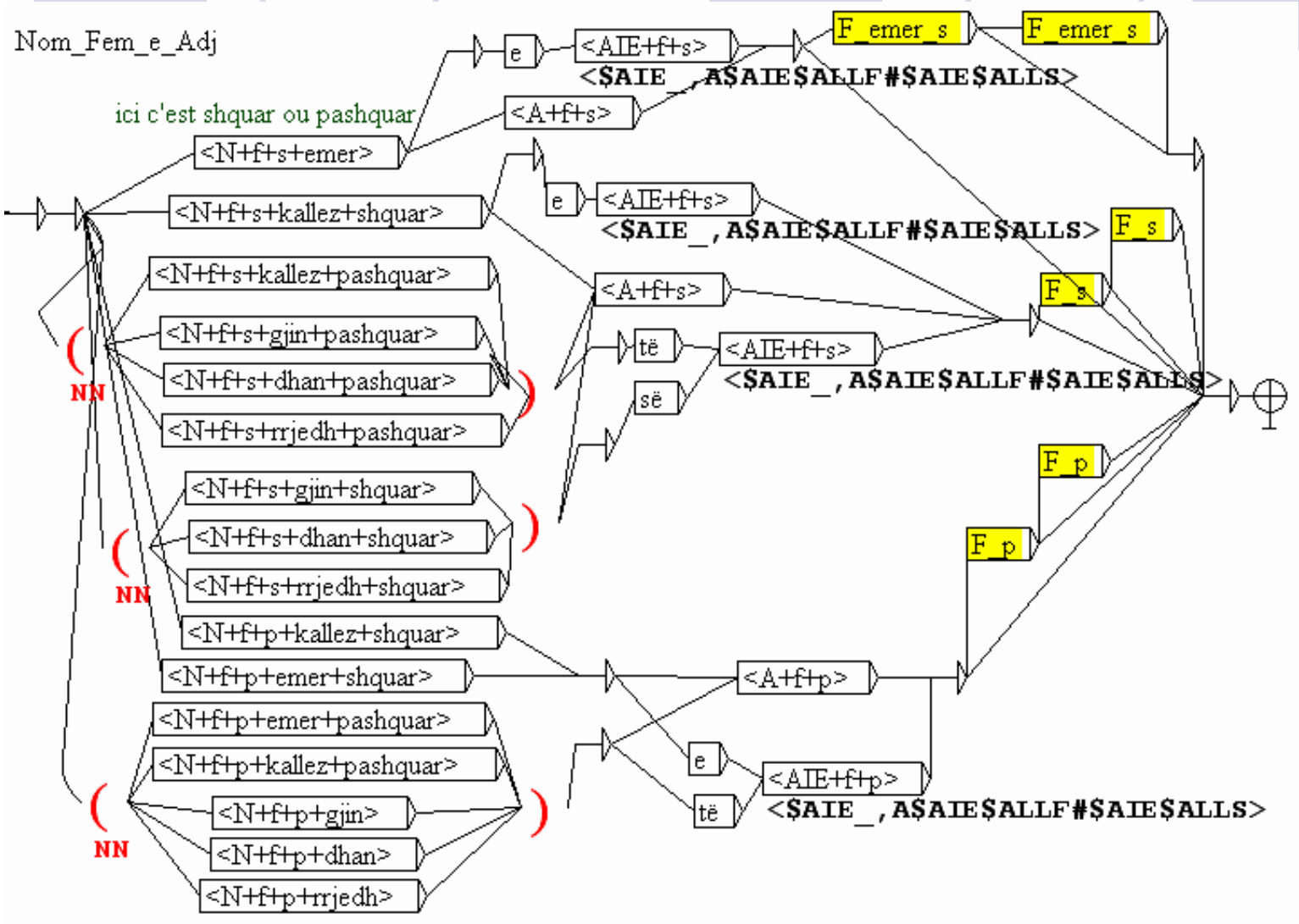
- **Grammar for the NP**



- **Two main graphs :**
  - **Nom_Masc_i_Adj for the masculine nouns**
  - **Nom_Fem_e_Adj for the feminine nouns**

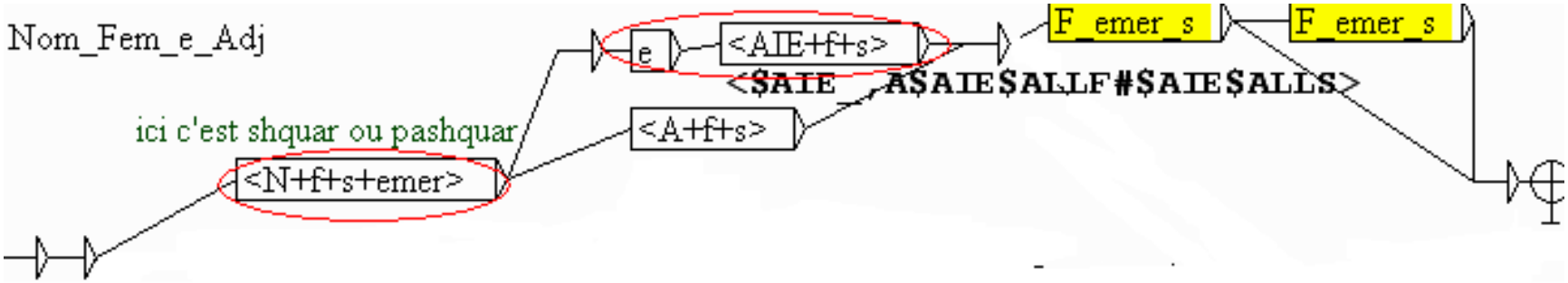
- **six subgraphs : three for the masculine and three for the feminine syntagms, depending on the declension**

Main graph for the feminine syntagm

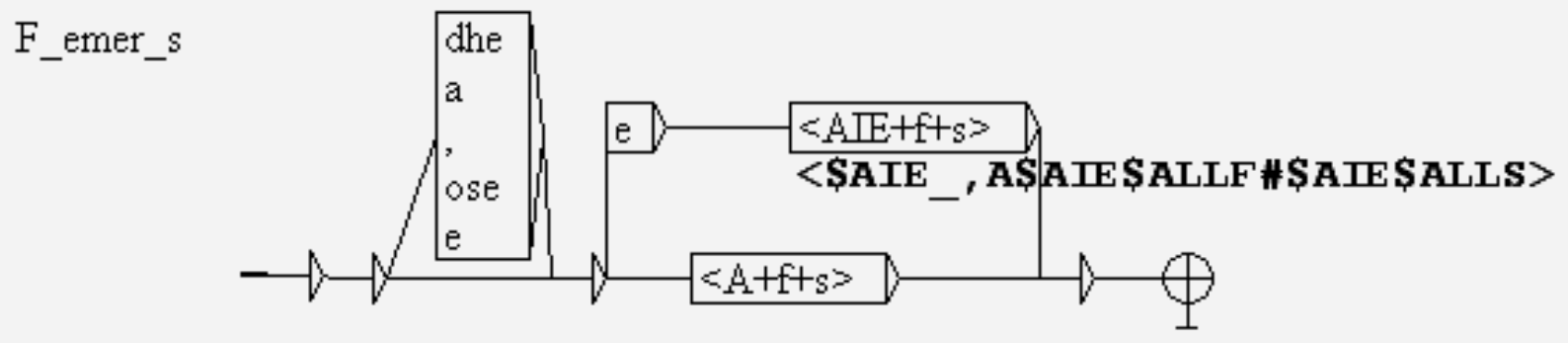


Exemple

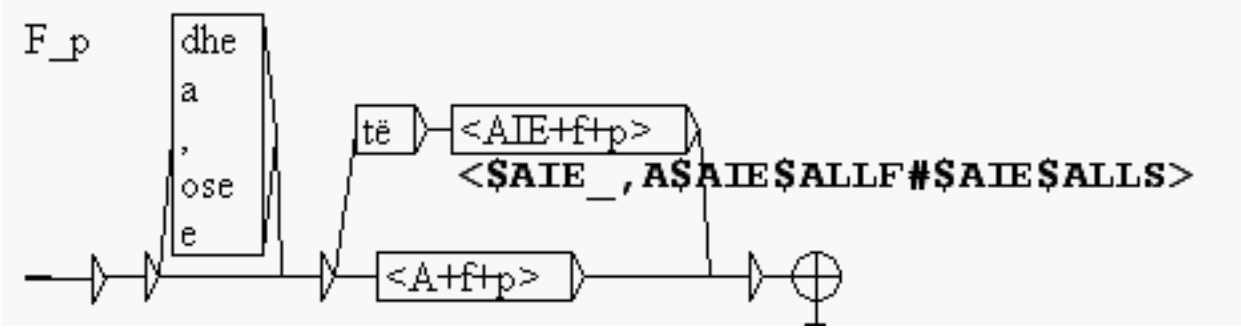
- **Concordance :**
- **bërthama e natyrshme dhe themelore**
- **Output :**
- **/<SN<natyrshëm,A+Genre=f+Nombre=s+FLX=Adj_ëm_ie+ie>>**
- **bërthama is a nominative feminine noun**
- **e natyrshme is the feminine form of i natyrshëm**
- **$AIE_ is the lemma natyrshëm**
- **<natyrshëm,A+Genre=f+Nombre=s+FLX=Adj_ëm_ie+ie>**



Example

- **dhe = and (no confusion with country because it is into the graph)**
- **themelore is a feminine adjective**
- 

_s
- The only difference between F_s and F_p is the number s or p



Results : 126 occurrences with 24 errors

Concordance for Text droits de l'homme.not

Reset | Display: 5 | characters / word forms | before, and 5 after. Display: Matches Outputs

Before	Seq.
inimi,	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
it dhe	kanë të drejtë/<SN<drejtë,A+Genre=f+Nombre=s+FLX=Adj_Ba+Ba+ie>>
enale	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
tës në	liri të plotë/<SN<plotë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
të lirë	dhe të barabartë/<SN<barabartë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
do të	jetë e nevojshme/<SN<nevojshëm,A+Genre=f+Nombre=s+FLX=Adj_ëm_ie+ie>>
het të	jenë të përgjithshme/<SN<përgjithshëm,A+Genre=f+Nombre=s+FLX=Adj_ëm_ie+ie>>
hkush	ka të drejtë të marrë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie><drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë të marrë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie><drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
hkush	ka të drejtë/<SN<drejtë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
shëm	dhe i vërtetë/<SN<vërtetë,A+Genre=m+Nombre=s+FLX=Adj_Ba+Ba+ie>>
jekur	kanë të drejtë/<SN<drejtë,A+Genre=f+Nombre=s+FLX=Adj_Ba+Ba+ie>>
etohet	si e drejtë/<SN<drejtë,A+Genre=f+Nombre=s+FLX=Adj_Ba+Ba+ie>>

The results : analyse of the errors

- **24 failures**
  - ex: ka të drejtë = he has the right to
    - but as a noun ka = an ox,
    - të drejtë = the right is like the adj. right
    - so, ka + adj is falsly « recognised » 14 times
  - Ex: kanë të drejtë = they have the right to
    - But as a noun kanë = a carafe
  - Ex: dhe të barabartë = and equal
    - As a noun dhe = a country (2 errors with dhe)
  - Ex: si e drejtë ; si = as, but si is also a musical note !

  - Ambiguities : midis = middle noun and prep
    - liri të plotë is recognised twice
    - liri të plotë = larger freedom or = the 'complete' linen

    the second one is false

102 successes: examples

- 3. Vullneti popullit është baza e pushtetit shtetëror; ky vullnet duhet të shprehet në zgjedhje periodike dhe të lira të cilat duhet të jenë të përgjithshme dhe votimi i barabartë, si dhe me votim të fshehtë ose sipas procedurës përkatëse të votimit të lirë.
- 1. Gjithkush ka të drejtën e shkollimit. Arsimi duhet të jetë falas, të paktën në shkollat fillore dhe të ulëta. Arsimi fillor është i detyrueshëm. Arsimi teknik dhe profesional duhet të zgjerohet e arsimi i lartë duhet t'u bëhet i mundshëm të gjithëve në bazë të aftësisë.

we remember :many other morph. graphs (Nooj 2007, NLDB 2007)

- **Open lists**
  - **For numeral numbers,**
  - **For ordinal numbers**
  - **For ‘XY’ words where X is a number**
  - **For ‘XY’ words where Y is a verb**
    - **And X is a number**
    - **And X is an prefix**

- **For imperative verbs (agglutination with a clitic)**

Conclusion

- The text has regular sentences, and lot of combinations don’t occur
- Some nouns are recognised by a grammar, because they include an hyphen
- Lots of other forms have to be included into the nominal graph :
- Pronouns
  - They can have a particle
  - They have a flection
- determiners must be included ;
  - Some are preposed and combined with the particle for the genitive
  - Lot of them are postposed
- Some forms are not very well established, so we must include some morphological graphs able to add or suppress mute e

- **we intend to make a grammar for the verbs**
- **the conjugation of verbs needs much more complicated graphs :**
- **The verb can have several components**
- **Clitics can be inside the compound verbal form**
- **These clitic forms are partly ambiguous**
- **Clitics can be agglutinated with the verb in the Imperative mode**
- **We have a good dictionary for verbs, but we have to do a syntactic and contextual grammar of the verbal syntagm.**