Proceedings of the
Nooj 2010
International Conference and Workshop

May 27, 28, 29 2010 Komotini Greece
Democritus University of Thrace

Edited by:

Zoe Gavriilidou
Elina Chadjipapa
Lena Papadopoulou
Max Silberztein

Table of contents

PREFACE 5

Learning the Greek language via Greeklish 9
Anastasia Georgiadou, Alex Karakos, John Papaioannou

The quiver of the algebraic mathematical models: The case of the electronic spell-checker 17
P. Kambaki –Vougioukli, T. Vougiouklis

A Corpus Based NooJ Module for Turkish 29
Mustafa Aksan, Ümit Mersinli

Arabic Compound Nouns processing: inflection and tokenization 40
Ines Boujelben, Slim Mesfar, Abdelmajid Ben Hamadou

Morphology based recognition of Greek verbs with Nooj 52
Angeliki Efthymiou, Zoe Gavriilidou

NooJ disambiguation local grammars for Arabic broken plurals 62
Samira Ellouze, Kais Haddar & Abdelhamid Abdelwahed

Greek neoclassical compounds and their automatic treatment with Nooj 73
Gavriilidou Zoe, Papadopoulou Lena

Deriving Adjectives and Nouns from Numerals 84
Kristina Vučković, Sara Librenjak, Zdravko Dovedan Han

Version 4 Greek NooJ Module: Adverbs, acronyms and words with Latin characters 95
Papadopoulou Lena, Chatzipapa Elina

Assignment of Character and Action Types in Folk Tales 102
Piroska Lendvai, Tamás Váradi, Sándor Darányi, Thierry Declerck

Named Entities in Chinese 112
Ying Yang, Miloš Utvić, Gordana Pavlović-Lažetić, Duško Vitas

Recognition of Libyan Person names using NooJ plateform 125
Abdelsalam Almarimi, Abdelmajid Ben Hamadou, Khaled Hussain, H. Fehri

Improved Parser for Simple Croatian Sentences 134
Kristina Vučković, Božo Bekavac, Zdravko Dovedan Han

Towards Parsing Croatian Complex Sentences: Dependent Noun Clauses 142
Vanja Štefanec, Kristina Vučković, Zdravko Dovedan Han

Vietnamese classifiers processing for nominal syntagms extraction 152
Hồ Đình Océane

Selection criteria for method of translation and some suggestions for the platform NooJ 159
Hajer Sahnoun and Kais Haddar

Proposal of a framework for the representation of Arabic named entities to use the transfer approach with NooJ 172
Hela Fehri, Kais Haddar and Abdelmajid Ben Hamadou

Greek Professional nouns processed with NooJ 183
Chadjipapa Elina, Papadopoulou Lena

Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform 192
Abdelmajid Ben Hamadou, Odile Piton and Héla Fehri

Recognition of negative paraphrases in Spanish 203
Angels Catena, Judith Sastre

Colour semantics and ambiguity, processing approaches with NooJ 212
Marcel Puig Portella

Automatic Transformational Analysis and Generation 221
Max Silberztein

Mary Astell's words in *A Serious Proposal to the Ladies* (part I), a lexicographic inquiry with NooJ 232
Hélène Pignot, Odile Piton

Building a Sanskrit module in NooJ: Basic resources 245
Vanja Štefanec

Greek in the age of corpora: Challenges and solutions 257
Dionysis Goutsos

A New Greek Corpus 270
Dimitra Alexandridou, Anna Anastassiadis-Symeonidis

PREFACE

The present volume includes 26 from among 46 papers that were presented at the 2010 Nooj Conference which was held at the Democritus University of Thrace (Komotini Greece) in May 27-29, 2010.

Nooj is a powerful corpus processor a) permitting sophisticated queries presented in form of concordances and b) supporting statistical analyses and information extraction. However, it is mainly a user friendly environment that allows a wide range of linguistic applications with the use of structured libraries of linguistic resources.

The construction of such libraries is facilitated by specific tools allowing the formalization of orthographic, morphological, lexical, syntactic and semantic linguistic data. To achieve that, Nooj consists of five different parsers: an orthographic parser, which is basically a flexible tokenizer, a morphological parser performing both derivational and inflectional analyses, a lexical parser for looking up vocabulary items (from morphems to discontinuous frozen expressions), a syntactic parser providing structured annotations and a semantic parser which is able to produce paraphrases automatically. These parsers are connected via a Text Annotation Structure (TAS) that stores not only results produced by each parser but also unsolved ambiguities.

The papers included in this volume present data concerning the spelling, morphology, syntax, lexicon and semantics of fifteen different languages (Greek, French, Hungarian, Arabic, Chinese, Catalan, Spanish, Vietnamese, Bulgarian, Croatian, Russian, Turkish, Armenian, Sanskrit, Serbian).

More precisely the orthographical level is the research domain of the joint papers of Georgiadou, Karakos & Papaioannou and that of Pinelopi Kambaki-Vougioukli & Th. Vougiouklis.

Morphology issues are studied in the papers of Aksan & Mersinli, of Boujelben, Mesfar & Ben Hamadou, of Efthymiou & Gavriilidou, of Ellouze, Haddar & Abdelwahed, of Gavriilidou & Papadopoulou, of Vučković, Librenjak, & Dovedan Han.

Mustafa Aksan and Ümit Mersinli in their paper "A Corpus Based NooJ Module for Turkish" present the design, implementation and testing processes of a corpus-driven Nooj module for morphological tagging of Turkish.

Ines Boujelben, Slim Mesfar & Abdelmajid Ben Hamadou in their paper "Arabic Compound Nouns Processing: Inflexion and tokenization" describe a new approach for Arabic Compound Nouns inflection and tokenization processing.

Angeliki Efthymiou and Zoe Gavriilidou in their paper "Morphology based recognition of Greek verbs with Nooj" study verbal derivation in Greek.

Samira Ellouze, Kais Haddar & Abdelhamid Abdelwahed in their paper "NooJ disambiguation local grammars for Arabic broken plurals" present a linguistic approach for the elimination of ambiguities between Arabic broken plural and the other grammatical categories.

Zoe Gavriilidou and Lena Papadopoulou in their paper "Greek neoclassical compounds and their treatment with Nooj" study Greek neoclassical compounds and show how Nooj morphological grammars facilitate the automatic recognition of such compounds.

Finally, in their paper "Deriving Nouns from Numerals" Kristina Vučković, Sara Librenjak and Zdravko Dovedan Han discuss formation of nouns and adjectives from numerals in Croatian language using NooJ morphological grammars.

The papers of Papadopoulou & Chadjipapa of Lendvai, Váradi, Darányi & Declerck, and of Ying Yang, Utvić, Pavlović-Lažetić & Vitas focus on the lexical level.

Elina Chadjipapa and Lena Papadopoulou in their paper "Version 4 Greek NooJ Module" describe the Nooj dictionaries built for the Greek module.

In their paper "Assignement of Character and action types in folk tales" Piroska Lendvai, Tamás Váradi, Sándor Darányi & Thierry Declerck develop NooJ lexicons and grammars in order to extract two types of content descriptors: characters and their actions.

Ying Yang, Gordana Pavlović-Lažetić Miloš Utvić and Duško Vitas in their research "Named entities in Chinese" built linguistic resources for NooJ, such as dictionaries to help locate named entities.

More papers focus on the syntactic level. Abdelsalam Almarimi, Abdelmajid Ben Hamadou, Khaled Hussain and Héla Fehri in their paper "Recognition of Names of Libyan Persons Using Nooj Platform" propose a system for the recognition of Libyan Person Names in order to translate them into English using the NooJ platform.

Kristina Vuckovic, Božo Bekavac and Zdravko Dovedan in their paper "Improved Parser for Simple Croatian Sentences" describe the improvements made for the Croatian syntactic parser.

In their paper "Towards Parsing Croatian Complex Sentences: Dependent Noun Clauses", Vanja Štefanec, Kristina Vučković and Zdravko Dovedan describe a model for partial parsing Croatian complex sentences.

Océane Ho-Dihn in her study "Vietnamese classifiers processing for nominal syntagms extraction" presents the implementation of Vietnamese classifiers in the nominal syntagm modelization.

In "Selection criteria for method of translation and some suggestions for the platform NooJ", Hajer Sahnoun and Kais Haddar focus on presenting a linguistic approach for automatic translation with Nooj.

Finally in “Proposal of a framework for the representation of Arabic named entities to use the transfer approach with NooJ” Héla Fehri, Kais Haddar and Abdelmajid Ben Hamadou propose a framework for the representation of Arabic Named entities based on the structure of features independently of lexical categories.

The paper of Elina Chatzipappa and Lena Papadopoulou “Greek Professional nouns processed with Nooj” focuses on the morphology and syntax interface.

The semantic level is studied in the works of Ben Hamadou, Piton & Fehri, of Catena & Sastre, of Puig and of Silberztein.

Addelmajid Ben Hamadou, Odile Piton and Héla Fehri in their paper “Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform” study the extraction of relation between Named Entities based on functional relations.

Angels Catena and Judith Sastre in their paper “Recognition of negative paraphrases in Spanish” propose an approach in order to identify negative paraphrases in Spanish in a question answering system.

In “Colour semantics and ambiguity, processing approaches with NooJ”, Marcel Puig offers solutions for the disambiguation of the semantic field of colours.

In his paper “Automatic Transformational Analysis with NooJ” Max Silberztein presents a new automatic transformational engine for NooJ, capable of producing all the paraphrases of any given sentence

The work “Les mots de Mary Astell, étude lexicologique, grammaticale et sémantique d'un texte du XVIIe siècle au moyen de la plateforme linguistique NooJ”, of Hélène Pignot and Odile Piton and the one of Vanjia Štefanec “Building a Sanskrit module in NooJ: Basic resources” both focus on the semantic, morphological and the orthographic level.

Finally the works of Dionisis Goutsos and of Alexandridou Dimitra and Anastasiadis-Symeonidis Anna focus on cospus linguistics.

Dionisis Goutsos, in his paper “Greek in the age of corpora: Challenges and solutions” offers a state-of-the-art description of corpus research on Greek, focusing on developments in corpus linguistics.

Alexandridou Dimitra and Anastasiadis-Symeonidis Anna in their joint paper “A New Greek Corpus” describe the construction characteristics and advantages of a new Greek corpus.

Learning the Greek language via Greeklish

Anastasia Georgiadou[(1),] Alex Karakos[(2)], John Papaioannou[(3)]

[(1)(2)(3)]*Department of Electrical and Computer Engineering*
Democritus University of Thrace

Abstract

Learning Greek as a second (L2) or foreign (FL) language has drawn theattentionof many researchers throughout time. There is a number of different ways to study a language, each of which has advantages and disadvantages. A dictionary is amongst the first things a foreign language student uses and is always a practical tool independently of the way one student might choose. Reading comprehension is significantly improved by the use of a dictionary, especially when this includes the way words are pronounced.

The aim of our proposal is the development of an assistance software for learning the Greek Language via Greeklish. Since, the basic vocabulary of a language is the basis of understanding the language itself, the dictionary proposed aims to make the basic Greek words easier to pronounce as well as to give the explanation of the word in English.

The programming language used towards the implementation of the software and the development of a user-friendly Graphic User Interface (GUI) is C++ combined with Win32 API. Moreover, the standalone application Greeklish Converter v1.00 implements the conversion of Greek to Greeklish characters.

The project forms a Greek to English and English to Greek dictionary that also provides the user with the pronunciation of the Greek word, typed in Latin characters (Greeklish).The algorithm of the project searches for a key word in the English or Greek word list, depending on the word the user is typing and the dictionary that is selected. More specifically, it implements a full-match or partial-match search, by doing character by character comparison and presenting the translated word that corresponds to the string of characters or words that the user typed, and matches it to the corresponding word in the word list of each dictionary. The process continues until there is a full-matching. After the word is located, the Greek word is translated to Greeklish through an external call to the executable program Greeklish Converter v1.00.

Greeklish Converter v1.00 is a standalone program that accomplishes the transliteration of a whole text written in Greek to Greeklish and vice versa. All its functionality is based on its ability to transliterate a single word. The algorithm that follows is based on reading and then fragmenting the input string. Every single part of the input string is characterized as a word. The word is being transliterated or not after a specific process.

The aim of this software is to provide a useful tool, a standalone software for each user that desires to learn the Greek language individually. Moreover, it aims to be involved, as an assistance tool, in the educational process for learning Greek as a second or foreign language.

1 Introduction

The term Greeklish refers to greek language written with the latin alphabet – either through transliteration or transcription. Greeklish can be used to applications such as e-mail, IRC and instant messaging, short message service and between Greek people living in other countries (Koutsogiannis, 2003 , Marinis, and al. 2007). This way of typing greek words is popular because it is easier and quicker to type and useful when there is no availability of Greek fonts (Tsourakis, and al. 2007). Converting documents from Greek to Greeklish and vice versa is being accomplished either with transliteration or transcription (Androutsopoulos, 2006).Transliterationis the mapping of one system of writing to another, word by word, or letter by letter (Karakos, 2003). Through transcription, the sounds of one language are depicted on the best matching script of another language. ISO 843 and ELOT 743 are two standards for transliteration in Greeklish, a one-to-one correspondence. Even though there are several standards, Greeklish texts may include spell varity. Some basic transliteration types are shown in the Table 1.

	η	**υ**	**ω**	**ου**	**θ**	**ξ**	**χ**	**ψ**
Phonetic Transliteration	i	i/u	O	u/ou	th	x/ks	ch/h/x	ps
Orthographic Transliteration	h/n	y/u	w/v	oy/ou	8/0	ks/3	x	ps
Keyboard Transliteration	h	Y	w/v	oy	u/q	j	x	c

Table1. Basic transliteration types

2 Learning foreign languages

Older approaches have suggested that there are three general ways approaching how foreign languages can be taught (*Nation, 2002*). The first one is to learn the general rules, learn and teach comprehension and vocabulary and understand the language's basic grammar structure. The second one is the integration of cultural differences .The third approach claims that it is important to master mother language, learn English as a second language and a third language by preference. As far as it concerns, learning Greek as a second/foreign language, its certification means good knowledge of grammar and writing and also vocabulary (*Hüllen, 2006*). According to the communicative approach language means interaction; it is an interpersonal activity and has a clear connection with society. Language study examines the use of language in context, both its linguistic context (what is said or written before and after a given piece of discourse) and its social, or situational, context (who is speaking, what their social roles are, why they have come together to speak)" (Berns, 1984).

While learning a foreign language it is very helpful to use dictionaries for word meanings and thesaurus. That helps with memorizing the language's vocabulary and also understanding and pronouncing the vocabulary (Holton, 2007). After all, the alphabet of several foreign languages, including English, is a Latin-based alphabet. Some Greek to English and English to Greek dictionaries provide the user with the pronunciation of the Greek word, typed in Latin characters, to make the word easier to pronounce (*Joseph, 2009*).

Greek language is considered to be one of the 40 major languages, according to Bernard's Comrie book, The World's Major Languages. After all, it is well known English

language was influenced by the Greek language. According to professor George Kanarakis there is a great interlanguage influence from one language to another when native speakers come in touch. The English language is beholden to Greek for a major part of its vocabulary. Greek has played a large part in the English language development.

3 Programming languages and tools

The programming language used towards the implementation of the software and development of a user-friendly Graphic User Interface (GUI) is C++ combined with Win32 API.Win32 API**,** the 32-bit API for Windows, isMicrosoft's core set of application programming interfaces (APIs) available in the Microsoft Windows operating systems and is designed to be used with C and C++. It helps an application program to interact with the operating system. Moreover, Microsoft Windows SDK provides tools and tutorial for creating software using WIN32 API.

The standalone application Greeklish Converter v1.00 implements the conversion of Greek to Greeklish characters. This application accomplishes the transliteration of a whole text written in Greek to Greeklish and vice versa and can be used as a standalone or an assistance program.All its functionality is based on its ability to transliterate a single word. C++ is used because it is compatible with Greeklish Converter, which was implemented in C++, is portable and has a very common compiler. That means that a C program can be compiled for a very wide variety of computer platforms and operating systems with little or no change to its source code.

The words lists used in software presented, can be found in Freelang open-source software.Freelang's dictionary is a freeware software for windows, very easy to install and includes two word files, the English to Greek words list, with 21.527 words and the Greek to English words list, with 21.237 words.

4 The algorithm

The algorithm used in our software issearching for the key word either in the Greek or in the English word list. A partial or full match of the word takes place, by matching characters and the translated word, which corresponds to the typed characters, is being displayed.The same process continues until full match takes place. When the word from English to Greek is located, then for each character of the word a search in a map container takes place.The algorithm's map includes transliteration rules that define the right Latin character matching each character of the Greek word. This process gives the Greeklish meaning.The stressed letter of the entry key-word is being displayed in bold and capital.In the case of the Greek-English dictionary the Greek to Latin characters matching,happens in real time. In the case user selects the key-word from the word list, the algorithm searches the meaning in the appropriate file. The Greeklish transliteration follows the same process as above.

The main advantage of the algorithm is that the Greeklish meaning of the word helps to easier understand the phonetic pronunciation, while the stress makes it easier to pronounce the word.

The dictionary that is supported from the algorithm above can be used either as a stand-alone software for each user that desires to learn the Greek language individually, or

as an assistance tool, in the educational process for learning Greek as a second or foreign language.

The executable's structure contains a folder named language which includes Freelang's word lists and a folder util which includes greeklish.exe (Picture 1.).

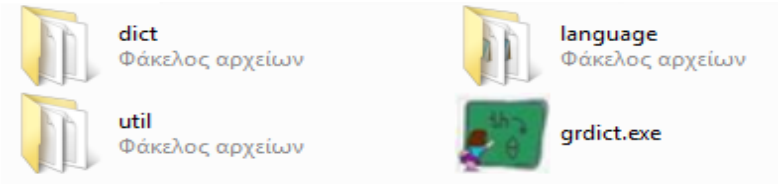


Figure 1. The executable's structure

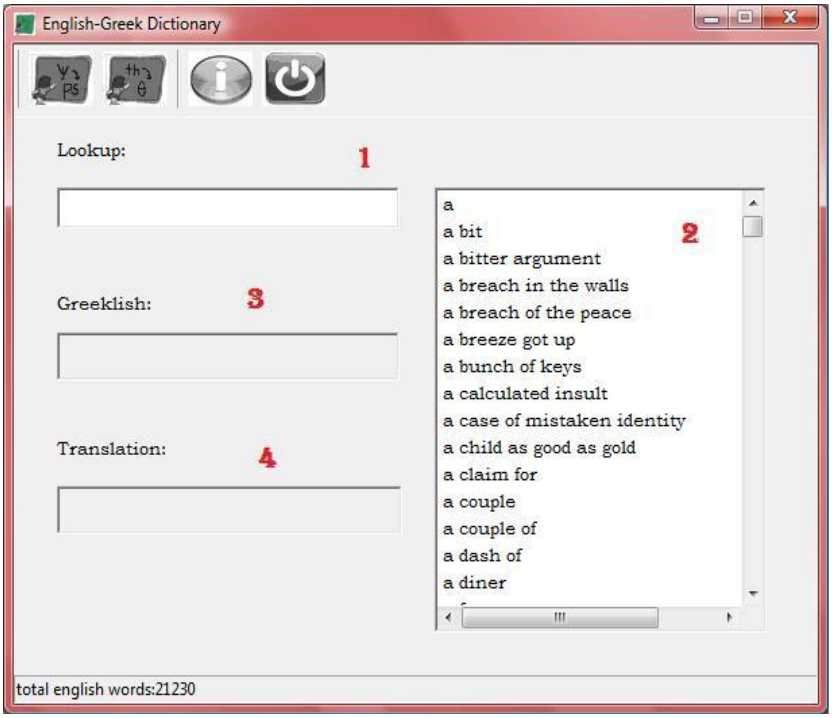


Figure 2. The initial screen

5 Dictionary interface

The initial screen that appears when executing the software (Picture 2.) displays the English-Greek Dictionary. The same screen appears by switching to the Greek-English Dictionary.

The application's toolbar consists of four buttons (Picture 3.). The two first buttons are used to switch between the two dictionaries. The third one displays an information window and the last one is the exit button.

Figure 3. Toolbarhot

While typinga word, in the lookup area, the list shown on the right part of the interface displays the list of the words that correspond to the typed letters (Picture 4.). The same happens with the Greek-English Dictionary.

Figure 4. Typing an English word

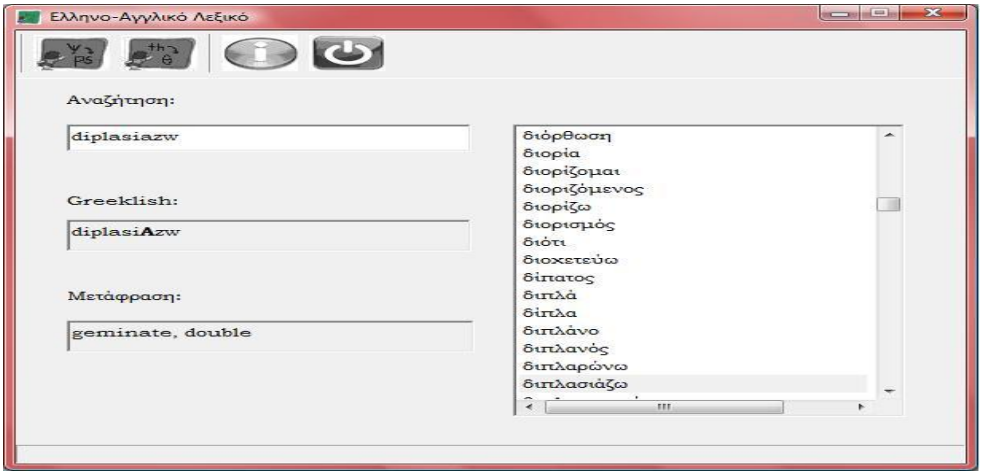


Figure 5. Typing an word in Greeklish

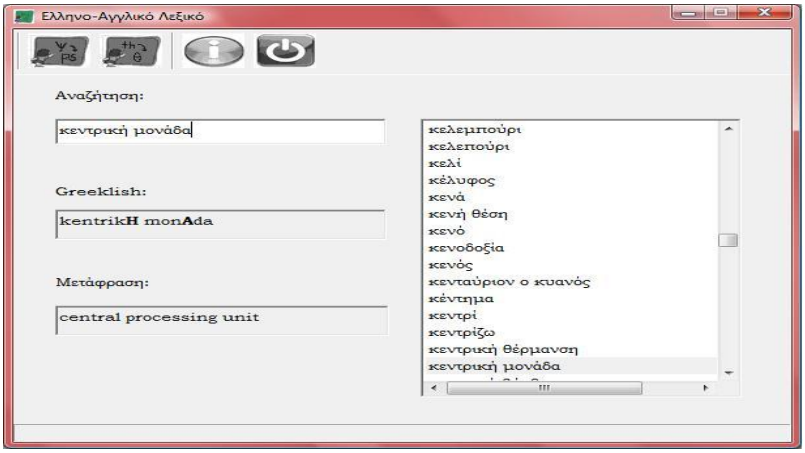


Figure 6. Intonation in Greeklish

Though the keyboard's language changes when the user switches from one dictionary to the other, he might accidentally type a Greek word with Latin characters. In Picture 5 it is shown that the Greeklish Converter transliterates the words in Greek and similarly as above, the same process takes place and the Greeklish and English translation of the typed word is being displayed.

In the Greeklish translation area of the interface (Picture 6.), the user can also see the word's intonation. The intonated letter is being shown in bold and capital. This is very helpful, since the user can understand how the word is pronounced, when it is also very important for paronymous words.

Figure 7. Example of an Italian-Greek Dictionary

6 Conclusions

The dictionary we present is a common general-use bilingual dictionary. Its main characteristic is that it can be used as assistance Software for learning the Greek Language via Greeklish. Each non-Greek user can understand the pronunciation of the words, because between the Greek and English meaning emerges the word typed in Latin characters.

The stressed letter being displayed in bold and capital makes the pronunciation more clear. The key-word search takes place either by selecting the word from the word list or by typing it in the appropriate field. The application is very easy and has a user-friendly interface. It can be used by individual user for business or communicating with Greek people when visiting Greece. It can also be used in the education process for understanding vocabulary, which is really important for learning a foreign language.

In the future, more words can be added in the already existing word list. There is the perspective of adding other languages' word lists except for English (as example in Picture 7. the Italian version), while the Greeklish transliteration is only connected with the Greek meaning. The possibility of using the application as a web service and also incorporating visual as far as phonetic intonation could be a prospective.

References

Androutsopoulos, J. 2006. 'Greeklish': Transliteration practice and discourse in a setting of computer-mediated Digraphia. Forthcoming in: Alexandra Georgakopoulou and Michael Silk (eds.) *Standard Languages and Language Standards: Greek, Past and Present*.

Berns, M. S. 1984. Functional approaches to language and language teaching: Another look. In S. Savignon & M. S. Berns (Eds.), *Initiatives in communicative language teaching.* Reading, PA: Addison-Wesley.

Freelang, http://www.freelang.net/

ISO 843. 1997. Information and documentation – Conversion of Greek characters into Latin characters. International Organization for Standardization, from http://www.iso.org.

Holton, D. 2007. 'Oi neoellinikes spoudes sto Panepistimio tou Cambridge', Institouto Neoellinikon Ereunon Ethnikou Idrymatos Ereunon, *Enimerotiko Deltio* 32 (December 2007), pp. 96-9.

Hüllen, W. 2006. 'Foreign language teaching – a modern building on historical foundations', I*nternational Journal of Applied Linguistics* Vol. 16 No. 1

Joseph, B. 2009. 'Why Greek is one of The World's Major Languages', Discussion Note, *Journal of Greek Linguistics 9*.

Karakos, A. 2003. Greeklish: An experimental interface for automatic transliteration . *Journal of the American Society for Information Science and Technology*, 54(11):1069-1074.

Koutsogiannis, D. & Mitsikopoulou, B. 2003. Greeklish and Greekness: Trends and Discourses of "Glocalness" . *JCMC* Vol 9 Issue1.

Marinis, T., Papangeli, A. & Tseliga, T. 2007. "Potizo" or "Potizw"? The influence of morphology in the processing of Roman-alphabeted Greek. In: Agathopoulou, E., Dimitrakopoulou, M. & Papadopoulou, D. (Eds). *Selected Papers in Theoretical and Applied Linguistics*, 17 International Symposium, English Department, Aristotle University. Thessaloniki: Monochromia, pp. 443-452.

Nation, P. 2002. 'Best practice in vocabulary Teaching and leraning', in Richards J. & W.Renandya (eds), *Methodology in Language Teaching*. An Anthology of Current Practice, Cambridge: Cambridge University Press, 267-272.

Sedgewick, R. 1999. "Algorithms in C++, Parts 1-4 (Fundamental Algorithms, Data Structures, Sorting, Searching)", 3rd Edition. Addison-Wesley, 1997. ISBN: 0-201-35088-2. Ενότητα 2.4.

Tsourakis, N. & Digalakis, V. 2007. "A generic methodology of converting transliterated text to phonetic strings case study: greeklish", In *INTERSPEECH*-2007, 1785-1788.

The quiver of the algebraic mathematical models: The case of the electronic spell-checker

P. Kambaki –Vougioukli[(1),] T. Vougiouklis[(2).]

[(1)]Department of Greek, Democritus University of Thrace, Panepistimioupolis, 691 00 Komotini, Greece, pekavou@helit.duth.gr
[(2)]School of Education, Democritus University of Thrace, Nea Hili, 681 00 Alexandroupolis, Greece, tvougiou@eled.duth.gr

Abstract

In this paper it is argued that even the so called simple mathematical models can and should be used both widely and consciously in various applied sciences so as to promote scientific development. In support of this view, there are presented certain applications of this specific theory in linguistics, mainly by the use of two general models introduced by Vougiouklis and Vougiouklis (2002). These are the dipole of product-quotient and the five-step model, which results from the general way of development of any branch of Mathematics. These different steps of development of any language are far from having being defined but undoubtedly their mastery may primarily lead to the better comprehension of the parameters and the potential of the structures so as to provide reliable conclusions. Therefore, when mathematical models are used in LT research extra attention should be paid so that every step should be investigated for a complete development of the model. In this paper we focus on the Cartesian product and quotient procedure and its applications in language teaching/learningor using. More specifically, it is suggested a better exploitation of an extremely useful tool, the electronic spell-checker with specific examples.

1 Introduction

Mathematicians trying to offer models to applied sciences other than Mathematics usually focus on a variety of aspects according to the outline defined by the specific science each time. Fairly enough, during this process, there were also introduced certain fields of Mathematics itself such as the *Category Theory* (MacLane, 1971). However, different sciences may ask from mathematics certain models by defining specific aspects. A quite extravagant realisation of such a demand is asking mathematicians to construct rather 'complicated and complex' models, as in Cryptography. Even more so it is to 'order' mathematical models for sciences with no connection with mathematics, at least at first sight. Then mathematicians 'create' mathematics as the *Fuzzy Theory*, the *Chaos Theory* orthe *Theory of Hyperstructures.* Asfor *linguistics,* it has always been associated with the use of mathematical models ever since it was first established as a science based on experiment and observation. This demand for models applicable in linguistic theory has been more of a rush during the second half of 20th century with Chomsky's "mathematicalization" of the language. However, this interaction between linguistics and mathematics is not new. Mathematics "have owed" to linguistics at least since Panini's times, possibly 5th century BC. As R. Mankiewicz (2000) mentions, if Greek mathematics

is based on philosophy, Indian mathematics is based on linguistics, and even more so on Panini's and the other great Indian linguists' work.

Contemporary linguistics and most of the so-called applied sciences borrow mathematical models mainly from Statistics; nevertheless there is the possibility that other branches of mathematics could provide the other sciences with useful models. We specifically refer to Algebra, that has supplied quite a few sciences with a good number of models enabling them in this way to organize themselves in a mathematical way.

Mathematical models have had quite a few applications in lexicography such as algorithms. Moreover, in our days with the expansion of computer science and its applications in electronic lexicography, the area of algebraic mathematical models seems to be a field which has a lot to offer towards the direction of a better organization of the science.

2 Previous Research

2.1 Models of mathematical models: Two General Models

In the creation of a mathematical theory several general or specific methods, are used. However, in order to have a specific theory considered complete, one is expected to work through several stages of process, or 'steps'. Of course, all these steps should not be expected to be of equal length or of equal difficulty. In Vougiouklis and Kambaki-Vougioukli (2000) there are suggested two general ways of development and study, applied in virtually every subject of mathematics. Actually, they are two procedures which are traced consciously or subconsciously, yet undeviatingly.

First General Model

We recall that for a 'complete' study in mathematics, virtually in all branches, one could identify the following steps:

(i) *The choice of the basic set of the study*

(ii) *Choice of the axioms, i.e. the basic rules of the construction*

(iii) *Construction*

(iv) *Morphisms. Principal mappings transferring the structures or basic constructing elements.*

(v) *Endomorphisms, i.e. transformations and their characteristic, invariant, elements.*

Second General Model

We believe that in Mathematics there are generally two inverse procedures:

(a) *the product, called Cartesian product, which is a very simple procedure and is based on the ordering of the objects, and*

(b) the quotient, which, by contrast to the product, is a very complicated procedure and not unique.

By following these steps in both general models, the exposition of a theory of Mathematics may be considered completed although more new constructions could be introduced and studied at every step.

The application of a certain structure as a model is an entirely different issue. Every applied science can occasionally use and -if not appropriate- reject mathematical models from every field of Mathematics; this by no means implies that the models are right or wrong but simply that they can be used or not for the specific purpose.

In this paper we propose a classification of the construction procedure and we point out some motivating examples from mathematics. Moreover, we claim that the proposed procedure does also exist in Linguistic Theory. The final implication is that some mathematical models may offer more than their creators intended to do.

Now let us try to elaborate on the five steps of the First General Model above:

In step (i), we have to specify the initial concepts and the set of elements to be studied. At this stage, it would be necessary to provide all possible elements to be used in every stage of application of the desired model.

In step (ii), we choose the appropriate axioms in order to build the structure, select the basic construction elements and establish the construction rules. These rules should be as limited - and appropriately selected - as possible so that they should not lead to inconsistencies, that is to say the destruction of the structure.

In the step (iii) of construction we form the structures and introduce new construction elements using proofs in every case. We test for identification of possible inconsistencies and if there are any, we return to step (ii) and redefine the axioms.

In step (iv), we define the morphisms which are the mappings transferring constructing elements from one structure to another or, more interesting, occasionally within a single structure. This transferring may reveal similarities in structures which, at first sight, might have seemed to be different. In other words, at this step we study the 'motion' of structures. At this step natural languages are ready to supply their users with the set of structures necessary to produce the '*Literature'* of each language.

Moreover, at this step, specific -but generally applied- mappings are investigated. An interesting example of these mappings are the *projections*, that is, any mapping f such that $f^2 = f$. Projections are extremely important in any study involving mathematical modeling as what a projection really expresses is that a mapping of a mapping actually is the initial mapping. The concept of the *parameter* also appears here and plays a crucial role.

The final step (v) focuses on morphisms in the same structure such as symmetry, reduction and projection which are usually called transformations. Invariant elements stay unchanged under mappings and this is of great importance in the process of structure construction. Furthermore, the invariant elements are sub-structures of the corresponding structures.

2.2 Quotient as a simple mathematical model

The Cartesian product is a very simple procedure and is based on the ordering. It can be applied in several cases of objects (grammar, syntax, lexis) or on more general classes as a general model applicable to every language (universal grammar). By contrast the quotient is

a very complicated procedure and not unique. With present paper we do not claim to introduce a new model but to emphasize on the fact that the two steps-Second General Model- have to be taken in order that the introduction of a new model should be considered complete.

Chomskyan Universal Grammar as a system of subtheories is actually a procedure of a product. N. Chomsky (1986) assumes that here the basic questions are the principles and parameters. Similarly, when U. Eco (1995) considers Latin and Vulgata appearing in Dante independent languages, then the pursuit of the perfect language is a Cartesian procedure.

Although it might appear to be *metalanguage*, we lay a procedure of quotient on the table. "Using a Cartesian product of subtheories, find an expanded theory; then, using a quotient, find a new theory which will actually contain the subtheories." (Vougiouklis et al, 2000, p.490)

The product of classes in partitions is quite widely used in the linguistic theory (see Gross, M., 1972).

Based on the respective theory form Mathematics (Vougiouklis,1994 and 1995), the following are suggested:*"…*in a given structure any arbitrary partition could potentially maintain certain axioms or related weaker axioms and it is in the researcher's hand to identify them*…"* (Vougiouklis et al, 2002, p.510).

If associativity (or commutativity) is valid, then, in a case of arbitrary partition, we obtain the so called *weakassociativity* (respectively *weakcommutativity)*. That is to say, there are class elements which connect these classes in some kind of associativity (respectively commutativity).

Here is an example from language, actually two partitions partially arbitrary:

(a) Consider the partition each class of which contains all possible synonym words. In this partition the majority of the classes of the words are *singleton*s, i.e. they consist of only one element, as the majority of the words have no synonyms. Yet, every partition in language is characterised by the synchronic occurrence of each item, as a word may have had a synonym in the past or may have one in future, but it has not any at present.

(b) Furthermore, if we refer to an electronic lexicon, e.g. spelling-check in a computer, then the number of the elements of the majority of classes is greater because they also include all possible morphological realizations of each item such as tense, person, gender, number, case, etc (also compounds and derivatives).

An example of class- behavior in the above partitions is the following:

In the first partition, word classes could possibly be as follows:

Actually thorough study offers security

actually *really* *in fact* as a matter of fact	*thorough* *complete* *detailed* *exhaustive*	*study* *research*	offers gives provides supplies	safety security

By choosing different representatives from each class, one could obtain a number which reaches 4x4x2x4x2 = 256 possible combinations. Of course not all of them are appropriate because they may be not in use or they mean something different. However, from a communicative point of view, they have a value as they could maintain communication, especially of the 'foreign-talk' type.

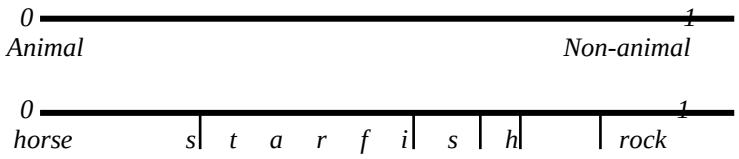
The same example becomes even more complex and complicated in table (b):

actually *really* *in fact* *as a matter of fact*	*thorough* *complete* *detailed* *exhaustive*	*study* *research* *studies* *researches*	*offers* *offered* *has offered* *had offered* *gives* *gave* *has given* *had given* *provides* *provided* *has provided* *had provided* *supplies* *supplied* *has supplied* *had supplied* *offer* *give* *provide* *supply* *have given* *have offered* *have supplied*	*security* *safety*

Here we have 4x4x4x23x2= 2.944 different possible combinations, not all of them plausible, of course**.**

2.3 The fuzzy sets

In 1965, Zadeh, a prominent professor of Electrical Engineering, inspired by Linguistic theory and more specifically Semantics and Pragmatics, first introduced the fuzzy sets in Applied Sciences with numerous applications in our everyday life including washing machines and air conditioning appliances. Briefly, Zadeh talks about the issue of criteria of membership and comes up with relations such as the one shown in the following graphical representation of *animal / non-animal continuum:*



Fuzzy sets theory could possibly offer more than one application in linguistic research (Joyce, 1976). In this paper we will present an application, namely the use of a bar instead of a Likert scale, application (4) below, which, we believe, might be useful to both teachers and researchers.

2.4 Electronic spelling correction procedure

The area of electronic lexicography seems to have potential of a good number of applications of algebraic models. More specifically, we believe that some of the so far mentioned small mathematical models could offer solid mathematical support to a lot of processes already used in electronic lexicography and, moreover, provide some innovative suggestions for experimentation. For example, M. Silberztein (2010) uses quotient procedure in order to discriminate affixes, morphemes and components as well as morphological variants and simple words. Similarly, M. Utvic, (2010), uses quotients in order to formulate rules for derivatives, while Chadjipapa et al (2010) also use small quotients in order to recognise proper nouns automatically. Clearly, the researchers are looking for quotients in order to organise the 'explosion' of possibilities which result from the productivity of the language and normally cause confusion and anarchy. Fairly enough, using an electronic tool for orthography implicates a good degree of risk, as Koutsogiannis et al (2002), point out. What we need is not only an effective tool but an equally effective guide to the teacher and/or to the user.

2.5 Some applications of the above models

In Kambaki-Vougioukli (2002a & b) and Kambaki-Vougioukli et al (2008) there are applications of the above mentioned models suggested, not necessarily for immediate use but mainly for consideration. Some of these applications are as follows:

(1) Economy of space in newspapers

Let us suppose that we have to handle the difficult problem of economy of space in a newspaper. It is a convention that a gap at the end of a 'word' manifests the end of the specific word. This manifestation may yield the implication that we are dealing with twenty-seven rather than twenty -six letters in the English alphabet, the twenty-seventh letter being a gap or an 'empty-space'. In mathematics, 'empty space' is symbolized by '0 (zero)' which is said to have first appeared as late as mid 300 AD. Such an ingenious use of gaps virtually leads to a quotient where we have as many subsets as the number of the letters consisting the longest possible word. Consequently, if we want to economize on space we should cut the '27^{th} letter', that is the gap at the end of each word, out. This would lead to strings of letters without gaps amongst them. This practice was quite common amongst Ancient Greeks who wrote without gaps between words maintaining in this way a better correspondence between spoken and written form, as Bauer (1988) points out. Would this ever happen, we should automatically encounter another problem: how would the end of each word be indicated? Should we possibly 'invent' a set of final letters? Yet, such a solution would be against the basic principle of economy of language. A more plausible solution would possibly be to use a set of letters which do exist, yet they are less

used and only in specific conditions, that is the list of capital letters. In this case our proposal could be formed as follows

> "*Abolish the empty space between words and indicate the beginning of each word using a capital letter"*

That is, in terms of our proposal:

> "*AbolishTheEmptySpaceBetweenWordsAndIndicateTheBeginningOfEachWordUsingACapitalLetter"*

At this point one should mention that experimentation of this kind is justified and has been quite often used in newspapers, magazines and advertisements. Another solution could be to use different colours, the most convenient being black and gray, in turns, or any other colours provided that it would be both plausible and economic.

(1) Quotients in partition of written documents:

As we have already mentioned quotients are absolutely arbitrary partitions. Although this fact is hardly acceptable by the vast majority of people, its certain applications are often considered to be self-evident. Let us examine such a 'self-evident' application concerning the partition of a written document. A certain partition of a written document into sentences, paragraphs and chapters involves the author's decision, who, apparently subconsciously, creates quotients whose contents refer to separate concept units. A further creation of arbitrary quotients takes place when s/he starts typing them. Now, the automatic change of line and page are undoubtedly not only arbitrary quotients but also temporary because they are bound to change with every adjustment of the top, bottom, left and right margins, line spacing, e.g. single spaced, double spaced etc, as well as the size and type of fonts to be used. Consequently, every set of words cannot and should not be characterised according to its content. Nevertheless, we do accept this fact to such extent that we do make chapter, page and line references, that is to say we accept the arbitrary as self-evident. Besides, the arbitrary characterises all languages according to Saussure and, before him, Aristotle. To conclude with this issue we would like to point out that alphabetical indexes are based on arbitrary layout.

(3) Quotients in rules and exceptions

It has already been stated that in the first step of the First General Model we lay down the basic set of the study. This procedure can be repeatedly applied, yet to a lesser extent, during the process of the study. Consequently, it is necessary to designate a specific set each time. When the multitude of the elements is small, then we could enter every single element separately. The vowels of the English alphabet are: A, E, I, O and U. Nevertheless, a full description of the set is mostly and usually the case. This description will be referred to as *the rule*. Studying the case of the rule, the following can be ascertained: If the rule expresses the full set, it is called an *absoluterule*; however, if it does not fully express the set, then the issue of the *exception*semerges. In
Kambaki-Vougioukli (2002c), a full description of every possibility is described.

(4) Fuzzy sets: Bar instead of scale

Questionnaires

One of the major tools used in an empirical research include *questionnaire*s, where Likert scales are normally and widely used. However, the escalation of a variable depends both on its nature and on the researcher's judgement. Such a decision is not easy to make in cases such as the compilation of questionnaires normally used in any research, including linguistic. Definitely there are certain more often scales, as the 5-grade Likert scale normally preferred to 3 or 4 grade scales.

0= I completely disagree, 1= I rather disagree, 2= I am somewhere between, 3 = I fairly agree, 4= I completely agree.

Such scales are characterised by certain elements-rules normally identified in every step. That is to say, they may either pinpoint a very positive beginning and a very negative end, or vice versa, as the above example. However, the most difficult part is the partition itself and where exactly the limits of the actual partition lie. The problem of discrimination of those categories is quite serious for the researcher but is even more so for the subjects of the research, who might need tedious explanations and, finally, they might miss the point of the research. In order to minimize such risks an alternative method based on the fuzzy theory introduced by Kambaki-Vougioukli et al (2008), suggests the use of a 'bar', whose two poles are defined by 0, on the left, and 1 on the right, as follows:

The participants, instead of the usual checking of one grade explicitly specified on the scale, they will have to 'cut' by a vertical line the continuum space at any point they think expresses best their answer to the specific question, as above.

Advantages of the suggested method include facilitation of even the least sophisticated subjects as anyone invited to answer will not need any special training or time-consuming explanations. This is because subjects will not have to discriminate the indistinct difference between two or more grades of a scale. To make our point more descriptive, we compare the use of a Likert scale to that of a bicycle or wheel chair going up or down a flight of stairs while the suggested 'bar' to a bicycle going up or down an inclined plane.

Finally, as far as dataprocessing is concerned, questionnaire-processing by using the bar gives the initiative to the researcher to 'escalate' the answers without having to decide in advance if s/he will finally need 3, 4,…,10 grades in order to be able to identify the parameters and clarify the differences between the grades. More important, s/he has the flexibility of establishing balanced or imbalanced scale according to the needs of the specific research.

One of the main characteristics of applied mathematics is the ability of different approaches, the ability of simplification of the form. That is why linear models are preferred and we tend to change continuum into discreet and vice-versa. Consequently the

bar offers the possibility of accurate processing which is the optimum for every researcher: from discrete into continuum and, even more, from single valued into multivalued or fuzzy.

3 Utilization of the small mathematical models in electronic spell-checker

Some specific aspects of the so far discussed small mathematical models could be identified in the process of checking the spelling of a word on the electronic spell-checker. Consequently, in this piece of research, a specific application of our two general models will be suggested, namely the procedure of checking the spelling electronically, an extremely widely used and useful process. We will try to provide examples as well as solid theoretical argumentation in support of our proposition.

However, before presenting the theoretical model, we should clarify certain issues concerning the process of checking the spelling of words in any language. What do we do when we check the spelling of unknown words?

First, let us focus on the issue of the *acoustic picture* of a written linguistic sign. In the majority of natural languages there is not normally a one - to- one correspondence, or *injection*, between a written and an oral sign. That is to say the representation of sounds in each word of a language is not standard or predictable and as a consequence there is *orthography* rather than *writing*. This is a problem known as the *historic orthography*, and it is due to the inflexibility of written language as compared to the flexibility of spoken language. This is one of the shortcomings users of a foreign language are normally confronted with but it could affect native speakers quite often, as well. So, when we try to transfer an acoustic picture into writing we have to use the actual alphabet of a language and write it down following the conventions put by the language. This is the so called *orthographic* word. However, in lexicology there is the distinction between the *orthographic* word and the *phonological word*, which is the representation of a linguistic sign not in the alphabet of the language but in phonetic symbols, usually those provided by the International Phonetic Association (IPA). This phonetic writing is nothing more than an attempt for a more accurate transfer of the actual pronunciation in writing, overcoming the obstacle of historic orthography.

In terms of mathematical models now, this transferring from one structure to another, this 'motion' of elements concerns step (iv) of our First General Model, that of the *morphisms*. The 'sound' of a word, its pronunciation, is actually a case of motion, an actual projection in writing of what we hear.

More specifically, in the process of writing, that is to say when we try to write down a word we hear, actually the acoustic picture of a linguistic sign, there is a sequence of projections taking place: (a) the first projection is that of the linguistic sign as pronounced by one of the interlocutors, the transmitter, and is received by the ear of the second interlocutor the receiver, (b) the second (projection) concerns the projection of what the receiver has caught into his/her mental lexicon in order to match the acoustic picture, the signifying, with a signified stored in his/her mental lexicon, and (c) the final projection takes place when the receiver writes down what s/he thinks the transmitter has pronounced. Fair enough, it is not always easy to conduct this projection successfully and there might be a number of elements inevitably lost in the process. Consequently, in all three the above mentioned projections, i.e. the orthographic and phonetic representations,

there are elements lost; nevertheless, it is clear there are more elements lost in the orthographic representation. This is because many pronunciation elements cannot be transferred in the standard orthography any more. This has been an inevitable and diachronic process in every language with tradition in written form.

Similarly and as far as the reading process is concerned now, there is a number of elements lost, such as etymological information or certain pragmatological elements including interlocution indicators and the author's actual intention (eg a threat/order interpreted in reading as a request, or vice-versa). Consequently, it is obvious that it is almost impossible to recover the real situation from a projection, as we cannot recover the actual house from an architectural plan.

In present piece of research, we put forward the problem of how to write correctly a word we hear, using the electronic spelling checker. More specifically, when trying to write on the computer a word we heard, there could be different possible ways of representation of almost each sound of the word in writing. What we would like to suggest is that an electronic spell-checker is expected to supply users with a long, complete list of suggestions. Such a process should involve writing down a literally exhaustive list of every possible representation of every sound in each word.

For example, let us take the word ***'lazy'***:

(a) The sound ***[ei]*** of the letter **<*a*>** could be possibly represented by a non-native speaker or even a less sophisticated native speaker, in the next five possible ways: ***a, ei, ai, ay, ae.***

(b)The sound ***[z]*** of the letter**<*z*>**could also be represented in four possible ways: ***z, s, zz, ss.***

(c) The sound ***[i]*** of the letter **<*y*>** could be also represented in five ways:***i, e, y, ie, ee.***

A visual representation:

a **ei** **lAzy** **ai** **ay** **ae**	**z** **s** **laZy** **zz** **ss**	**y** **i** **lazY** **e** **ie** **ee**

Overall, the total of the 'wrong' and 'correct' representations of the word 'lazy' comes to 5x4x5=100.

Consequently, when any of the above mentioned possibilities is entered, it is expected from the spell-checker to supply the user with a list of suggestions containing the word 'lazy'.

4 Suggestions for future research

Both of the suggested General Models could be applied wherever we have to take 'simple', short steps: the bipole product-quotient, i.e. the Second General Model, or/and certain steps such as the use of invariant from the First General Model.

(a) We hold that in this way the learners can be trained to group, to line things up, to express in a uniform way small language problems. That is to say, to be able to recognize certain language procedures they have already subconsciously mastered, or in other words to be language aware.

(b) Learners should also be encouraged to master the way of discovering the above mentioned 'simple' or small models, in such a way that they would be ready any moment to extract structures or rules in order to facilitate their own learning process. These structures needn't be learned by heart but learners should be taught how to reach them any moment, unless, of course, they themselves decide they want to memorise a specific structure. With this proposal we want to indicate that in the teaching procedure we should not provide the learners with numerous sets of prefabricated rules to be memorised and never used. By contrast we insist that learners should master the actual procedure of extracting the rules themselves, when needed, i.e. inductively.

(c) An idea for a project could be put forward, concerning the exploitation of every possible representation on the electronic spell-checker.

In terms of small mathematical models, this is a case of the Second General Model, the bipole quotient-product, a process, usually the quotient, is usually identified in rules. Only, in this case, we are actually seeking for products rather than quotients, as we need every possible representation of an acoustic picture of a linguistic sign, a word, in writing. The crucial importance of such a process can be identified on each individual computer, when the user enters any word and expects to find his/her choice amongst the alternatives offered by the spelling checker. One should also take into account the fact that in certain languages, such as Greek, the existence of the stress - mark multiplies the possibilities. Last but not least, the frequency of the suggested alternatives is an important issue and implies a lot of work and thought for languages which lack frequency lists such as Greek. Therefore, a useful and rewarding project could be an exhaustive list of any possible representations of each word in any language. Such a project would offer a lot to non-native and native-speakers of a language not only because it would save them time but also as it would boost their confidence in services offered by the specific tool.

References

Bauer, L. 1988. *Introducing Linguistic Morphology.* Edinburgh University Press, 8-10.

Chadjipapa, E., Papadopoulou, E., Gavriilidou, Z. 2010.New data in the Greek Nooj module: Compounds and proper nouns. *Procceedings 7th NOOJ,* 93-100.Cambridge Scholar Publishing, UK.

Chomsky, N.1986. *Barriers*. Boston, Mass: MIT Press.

Cross, M., Moscardini, A O. 1985. *Learning the art of mathematical modelling*. J.Wiley.

Eco, U. 1995. *Η αναζήτηση της Τέλειας Γλώσσας*. Αθήνα: Ελληνικά Γράμματα.

Gross, M. *Mathematical models in linguistics.1972.* Englewood Cliffs, NJ: Prentice Hall.

Joyce, J., 1976. Fuzzy Sets and the Study of Linguistics. *Pacific Coast Philology,* 11,39-42.

Kambaki-Vougioukli, P. 2002(a). *Στο πρότυπο των προτύπων της γλώσσας*. *Studies in Greek Linguistics 22,* 51-59. Thessaloniki.

Kambaki-Vougioukli, P. 2002(b). Μαθηματικά μοντέλα στη γλωσσική διδασκαλία. *Η διδασκαλία της ΝΕ ως μητρικής γλώσσας,* 75-87. ΔΠΘ, Κομοτηνή.

Kambaki-Vougioukli, P. 2002(c). Κανόνας και εξαίρεση. *Πρακτικά. 2ου Διεθνούς Συνεδρίου για τη Διδασκαλία της Νέας Ελληνικής Γλώσσας*, 193-200. Αθήνα.

Kambaki-Vougioukli, P., Vougiouklis, T. 2008. Bar instead of scale. *RatioSociologica, 3,* 49-56.

Koutsogiannis, D., Manouilidou, Ch. 2002. Ο αυτόματος ορθογραφικός έλεγχος ως μέρος των ηλεκτρονικών περιβαλλόντων παραγωγής και διδασκαλίας του γραπτού λόγου. Studies in Greek Linguistics 22, 91-100. Thessaloniki.

McLane, S. 1971. *Categories for Working Mathematicians.* Springer -. Verlag, Berlin.

Mankiewicz, R., 2000. The Story of Mathematics, Cassell & Co, Wellington House, London.

Silberztein, Max. 2010. Disambiguation tools for Nooj. *Procceedings 7th NOOJ,* 1-14. Cambridge Scholar Publishing, UK.

Utvic, Milos. 2010. The regular derivation in Serbian: principles and classification using Nooj. *Procceedings 7th NOOJ,* 148-157. Cambridge Scholar Publishing, UK.

Vougiouklis, T. 1994. *Hyperstructures and their Representations,* Hadronic Press. USA

Vougiouklis, T. 1995. Some remarks on Hyperstructures, Contemporary Mathematics, *American Math. Society,* Vol.184, 427-431.

Vougiouklis,T., Kambaki-Vougioukli P. 2000. On the Mathematics of the Language. *Procceedings 2nd Panhellenic Congress "New Technologies for Society and Culture"*, 486-491.

Vougiouklis, T., Kambaki-Vougioukli, P., 2002. Simple Mathematical Models in Language, (in Greek). *Proceedings of the 19th Panhellenic Congress of Mathematical Education,* Greek Mathematical Society, 506-515.

Zadeh, L.1965. Fuzzy sets. *Information and Control,* 12/2, 94-102.

A Corpus Based NooJ Module for Turkish

Mustafa Aksan[(1),] Ümit Mersinli[(2)]

[(1)(2)],*Mersin University*
Mersin, Turkiye

Abstract

This paper presents the design, implementation and testing processes of a corpus-driven Nooj module for morphological tagging of Turkish. It also underlines the morphological challenges specific to Turkish. Modeling and tagging processes involves both inflectional and derivational paradigms of present-day Turkish. Inflection of multi-word units and syntactic disambiguation are beyond the scope of the study.

1 Introduction

Beginning with Hankamer's *keçi* system (Hankamer, 1989), studies on the morphological analysis and tagging of Turkish have a relatively short history. Despite the progress made in computational analysis of Turkish, significant challenges remain that the agglutinative nature of the language brings into foreground. Among the rule-based, non-stochastic, root-driven, left-to-right processing approaches with the purpose above, we can mention Oflazer (1994b), Çiçekli (1997) for two-level formalisms and Akın (2007) for a letter-based approach. Specifically, Bayraktar (2008) and Bisazza (2009) are studies using NooJ environment as a graph-based corpus processor.

This paper presents a NooJ module for Turkish which aims to analyze and annotate derivational and inflectional affixes of Turkish and to assign lexical categories to the base forms.

The paper is organized in the following order. After the introductory remarks, Section 2 states the data source for the module. Section 3 exemplifies challenges in the morphological analysis of Turkish. Section 4 provides the overall architecture of the module. Section 5 presents the content and structure of dictionaries and the dictionary compilation process. Section 6 is devoted to the modeling of Turkish morphotactics. In the conclusion, we present plans for future releases of the module.

2 Data

Data of the study are derived from the ongoing Turkish National Corpus (TNC) Project[1] held at Mersin University, Turkiye.

3 Challenges

Turkish affixation:

[1] Further information on Turkish National Corpus is available through www.tnc.org.tr

i. has potentially unlimited[2] combinations as in (1)

(1) çözümleyiciliklerindenmişçesine
as if it is because of their analyticalness

ii. includes homophonous roots, affixes, buffer phonemes or their combinations as in (2), (3), (4)

(2) **yazın** sıcağı
heat (weather) ***of the summer***
senin **yazın**
your essay/handwriting/summer
yazın taraması
literature *overview*
adınızı **yazın**
write down your name
yazın git
go (there) ***in the summer***

(3) ev**i** aldı (Accusative)
bought ***the*** *house*
onun ev**i** (Possessive)
his *house*
benim ev**i**m (Buffer phoneme)
my house
ev imiş (copula *i*)
it had ***been*** *a house*

(4) Biliyorum **ki** komşuda**ki** baklava bizim**ki** gibi.
I know ***that*** *the baklavaki* ***(which is)*** *at the neighboring (country) is like* ***ours****.*

iii. lets recursive concatenations as in (5)

(5) duvarı yık+**tır**+**t+tır**.
cause *someone to* ***make*** *some other one to* ***get*** *someone to tear down the wall.*

iv. has homophonous forms which are included both in derivational and inflectional paradigms as in (6), (7).

(6) yap**ma** bebek gibi adjective-forming
like a baby doll
sakın dondur**ma**! negative
don't freeze (it)!
kaymaklı dondur**ma** noun-forming

[2] Güngör (2003) provides statistical information on Turkish affixation.

dairy ice-cream
dondur**ma**ya başladı gerund
began freezing

(7) dondur**ma**lı / dondur**malı**
*with ice-cream / **must** freeze*

dondur**ma**dan / dondur**madan**
*from the ice-cream / **without** freezing*

4 Architecture

The module is constructed with a *.nof* grammar (phonology_TR.nof) which is employed for in-root phonemic alternations, and a *.nom* grammar (morphology_TR.nom) including derivational and inflectional affixes of Turkish. Turkish lexicon is represented by four pre-tagged dictionaries (content_TR.nod, function_TR.nod, multiword_TR.nod, proper_TR.nod).

5 Dictionaries

Dictionaries are formed by the tokenization of a subcorpus of over 3 million words from TNC. Lemmatization of the tokens was done manually in cooperation with the members of Linguistics Department at Mersin University.
Dictionaries have the number of entries listed in Table 1.
The growing number of entries in the dictionaries are provided in synchrony with the texts imported to TNC Project and represents the lexicon of present-day Turkish.

	content_TR	**function_TR**	**multiword_TR**	**names_TR**
Verbs	1,269	-	9,112	-
Nouns	16,171	-	7,938	-
Onomatopoeia	528	-	213	-
Adjectives	2,238	-	492	-
Pronouns	33	-	38	-
Numericals	112	-	53	-
Adverbs	267	-	684	-
Conjunctions	-	61	69	-
Interjections	-	2	238	-
Postpositions	-	13	-	-
Proper nouns	-	-	-	35,152
Abbreviations	-	-	-	1,430

Table 1. Contents of dictionaries

The entire lexicon is split into four in order to reduce the number of ambiguities by assigning priorities to each dictionary as in Table 2.

content_TR.nod	L1
function_TR.nod	H2
multiword_TR.nod	H3
names_TR.nod	H1

Table 2. Priorities of dictionaries

Linguistic information included in the dictionaries are presented and exemplified in Table 3.

Lemma	**,POS**	**+alternation**	**+syllables**	**+phonology**	**+derivation**	**+inflection**
tat	,NN	soften_t	+1	+tı	+lı	
tat	,VB	soften_t	+1	+tı		+ar
af	,NN	double	+1	+tı		
kon	,VB		+1	+du	+ık	+ar

Table 3. Structure of dictionaries

In addition to the fields simplified and presented above, also tags such as <+end_V> "ends with a vowel", or <+end_l> "ends with consonant (L)" are added to related entries to be used as constraints in the morphological graph.

5.1 Lexical Categories

Lexical categories or part-of-speech (POS) tags in Table 4, mostly identified and tested by structurally distinctive features, are assigned to lemmas in the dictionaries. Problematic Noun/Adjective or Conjunction/Interjection distinctions were done through a number of morphological, distributional tests. Etymological or diachronic information is excluded from the study and non-productive affixes or root forms that do not exist in the present-day Turkish are left beyond the scope of pre-tagging. Affirmative particle or clitic +MI, although not a lexical category, is added to the tagset for practical purposes. Onomatopoeic words are considered as belonging to a separate category since they have their own derivational constraints.

Tag	**Lexical Category**	**Examples**
<VB>	Verb	*git, gel, dur, bak, kal, sus, gör, dök*
<NN>	Noun	*gece, hava, renk, fark, dost, oyun*
<PN>	Pronoun	*bu, kendi, hepsi, herkes, kim, öteki*
<NB>	Number	*iki,üç, beş, sekiz*

<AJ>	Adjective	*mavi, yeni, düz, dürüst, zeki*
<AV>	Adverb	*acaba, asla, bazen*
<PP>	Postposition	*gibi, göre, için, kadar, karşı, rağmen*
<ITJ>	Interjection	*aferin, sağol, haydi, hoşçakal, lütfen*
<CJ>	Conjunction	*ama, çünkü, meğer, üstelik*
<ON>	Onomatopoeia	*takır, vızıl, gürül*
<NP>	Proper Noun	*Atatürk, Mersin, Ümit*
<AB>	Abbreviation	*TBMM, TDK*
<MI>	Affirmative	*mi, mı, mu, mü*

Table 4. Lexical Categories

5.2 Phonemic Alternations

Turkish has in-root morphophonemic alternations mostly forced by vowel and consonant harmony rules. These variants as in the examples in Table 5 are included in the compilation of dictionaries with the rules stated below. The file *phonology_TR.nof* covers all in-root morphophonemic alternations of Turkish lexicon except for multi-word units.
Some rules such as *soften_double* are applicable to a small number of words as mentioned in the examples column of Table 5.
All other external morphophonemic phenomena like the addition of "n" to pronouns ending in a vowel as in "hepsi**n**de" *in all of them* or "bu**n**dan" *from this one* are processed within the morphological graph.

Alternation	**Rule**	**Examples**
double	<D>	af > affı zam > zamma
drop	<L><B><R>	akıl > aklını fikir > fikrimin vakit > vaktinde
dropsoften1	<B2>b	kayıp > kaybına kutup > kutbuna
dropsoften2	<B2>d	kayıt > kaydına nakit > nakde
dropsoften3	<B2>c	avuç > avcuna kutup > kutbuna
compound1	<B>	anaokulu > anaokulları
compound2	<B2>	elyazısı > elyazıları başağrısı > başağrıları
compound3	<B2>ç	ipucu > ipuçları

compound4	<B2>k	ayçiçeği > ayçiçekleri
compound5	<B2>ul	sultanoğlu > sultanoğulları
compound6	<B2>p	elkitabı > elkitapları
compound7	<B2>t	kesekağıdı > kesekağıtları
soften_ch	<B>c	ağaç > ağacı süreç > süreci
soften_k	<B>ğ	emek > emeği diyalog > diyaloğu
soften_g	<B>g	renk > rengi
soften_p	<B>b	kitap > kitabı mektup > mektubu
soften_t	<B>d	cilt > cilde dört > dördünü
softenpdouble	<B>b<D>	tıp > tıbbın muhip > muhibbi
softentdouble	<B>d<D>	zıt > zıddı
change_an	<B2>an	ben > bana sen > sana
change_m	<B>m	saklan > saklambaç dolan > dolambaç

Table 5. In-root phonemic alternations

5.3 Other lexical information

Constraints governing Turkish morphology, specifically in derivational processes, are not yet fully described and particular inflectional pecularities do not have adequate justifications. Thus, instead of adding semantic subclasses for lexical categories – at least for the current release - we have employed morphological tags denoting exceptional inflections such as Aorist "+Ar" or tags denoting derivational constraints similar in form to the related affix, such as "+la".

Detailed analysis of Turkish affixation can be further found in Underhill (1976), Kornfilt (1997), Lewis (1967) and Göksel (2005). On derivational constraints, this paper refers to Uzun et al (1992), and Uzun (1993),(2008).

6 Morphological graph

Morphological graph of the module is constructed as in Figure 1 – prefix 'FLX' denoting inflectional subgraphs.

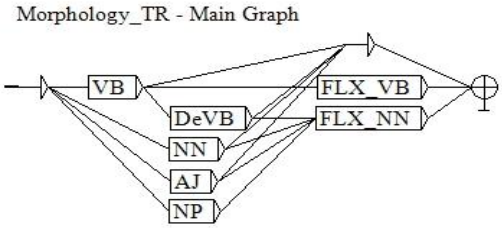


Figure 1. Morphology_TR.nom

6.1 Derivation

Considering the recursive nature of Turkish derivational affixation, related subgraphs are formed as in Figure 2. The first slot/level includes constraints and the second is recursive in itself.

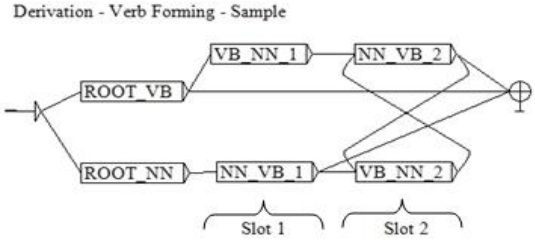


Figure 2. Sample derivational sub-graph

6.2 Inflection

6.2.1 Nominal Paradigm

In order to eliminate artificial ambiguities, mostly in the nominal paradigm, we have modeled nominal inflection in two separate subgraphs– one for base forms ending in a vowel and the other, with a consonant – as in Figure 3.

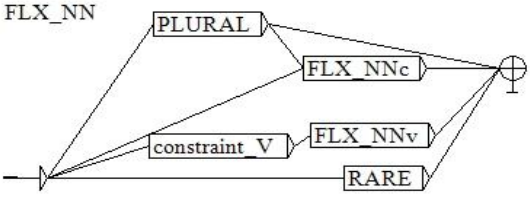


Figure 3. Nominal Inflection – Main graph

Külekçi (2001), Makedonski (2005), Oflazer et. al. (1994a), and Oflazer (1994b) are other studies which include models of Turkish nominal inflection.

6.2.2 Verbal Paradigm

Modeling of Turkish verbal inflection was less problematic when compared to nominals. Based on theoretical considerations discussed in Sebüktekin (1974) and Sezer (2001), verbal inflection subgraph is formed as in Figure 4.

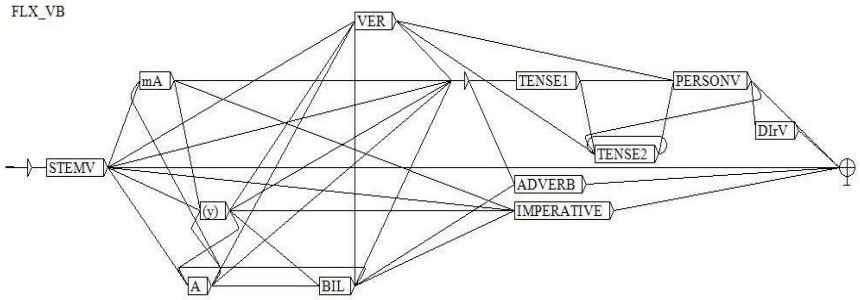


Figure 4. Turkish Verbal Inflection

7 Usage

7. 1. Annotations

(8), (9), (10) are sample annotations from the current release of NooJ Turkish module.

(8) çözümleyiciliklerindenmişçesine
çöz,VB
+(I)m_NN+lA_VB
+(y)+IcI_NN+lIk_NN+lAr
+I+n+DAn[ABL]+mIş[Per]
+cAsInA_AV

(9) biliyorum ki komşudaki durum bizimki gibi
bile,VB+yor+(I)m[1Psn]
bil,VB+(I)+yor+(I)m[1Psn]

ki,CJ

komşu,NN+DA[LOC]+ki[AJ]

dur,VB+(I)m_NN
duru,AJ+m[Poss]

biz,PN+(I)m[1Psn]+ki_PN

gibi,PP

(10) yıktırttır

yık,VB+DIr_VB+t_VB+DIr_VB

(11) yazın

yaz,VB+I_NN+n	*your handwriting*
yaz,VB+ın_NN	*literature*
yaz,VB+In(Iz)[IMP]	*write (imperative)*
yaz,NN+In_AV	*in the summer*
yaz,NN+I+n	*your summer*

7. 2. Concordances

(12), (13), (14) are sample queries and concordance lines.

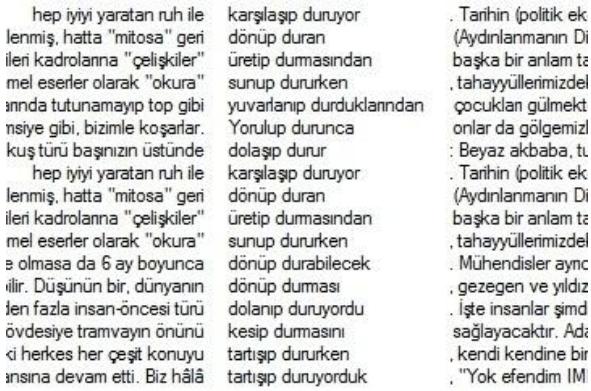
(12) <oku,VB>

nın sunduğu yöntemlerle	okuyor	. Kullanım değeri'nin yerine ç
r muyuz "Raskolnikov"u	okurken	? Kurdurmak zorunda mıyız?
ın kısaltılmış versiyonunu	okumayı	kendine hakaret olarak algıl
ıakaret olarak algılayan "	okur	"un klasikten ne umduğunu,
da olur yakından işitirdik.	Okuma	yazması olmayan biriydi. Ne
ir destan. Eve gidip iyice	okumak	ve anlamak isterdim. Okudu
ıak ve anlamak isterdim.	Okuduğumda	bir türlü pazar yerindeki o
RS-500, elektronik kitap	okumanızı	, RSS haberlerini takip etmer
; biçiminde uygulandığını	okuduklarımızdan	öğreniyoruz. Foucault'un "H
da olur yakından işitirdik.	Okuma	yazması olmayan biriydi. Ne
ir destan. Eve gidip iyice	okumak	ve anlamak isterdim. Okudu
ıak ve anlamak isterdim.	Okuduğumda	bir türlü pazar yerindeki o
nın sunduğu yöntemlerle	okuyor	. Kullanım değeri'nin yerine ç
r muyuz "Raskolnikov"u	okurken	? Kurdurmak zorunda mıyız?
ın kısaltılmış versiyonunu	okumayı	kendine hakaret olarak algıl
ıakaret olarak algılayan "	okur	"un klasikten ne umduğunu,
dele bir parça Osmanlıca	okuyabilen	varsa tanıdık çevrede, kesir

(13) <VB+r[Aor]><VB+mA+z[Aor]>

"kültürsüz kumazlığın" eline	geçer geçmez	bir şekilde "toplumsa
Dışarı çıkınca yola adımınızı	atar atmaz	cenk başlıyor. Çocuk
"kültürsüz kumazlığın" eline	geçer geçmez	bir şekilde "toplumsa
ı. Devinim ağırlıklı bir anlayış	ister istemez	beden yapının denge
adamları insanların bir yanlışı	bulur bulmaz	gülmediklerine, anca
e bir ekonomik kriz ortamına	girer girmez	toplumdaki "günah k
İnceleme heyeti'nden onay	alır almaz	donör araştırmalarına
rkezleri baskılanır ve bebek	doğar doğmaz	ağlayamaz. Bu da ol
adar kaçınılması ve bebeğin	doğar doğmaz	annesinin sütü ile be
ibi, tek eşliler çiftleşme sona	erer ermez	yeni bir eşin peşine d
gezegenle ilgili bir çevredir.	Doğar doğmaz	böyle bir çevreyle ka
bir çevreyle karşı karşıyayız.	Doğar doğmaz	, belki bütün organizr
apılan fizyonun keşfedildiğini	duyar duymaz	, bu elementin zincir r
ıastanın gözlerinde bu isteği	sezer sezmez	onlardan önce davra
duğuna göre, promosyon da	ister istemez	hekimlere yöneliyor !
ı da karısı Klytemnestra eve	döner dönmez	banyoda şişleyerek ö
klı siteye bulaşmış durumda.	Bulaşır bulaşmaz	sistemi çökerten ve l

(14) <VB+Ip[AV]><dur>



hep iyiyi yaratan ruh ile	karşılaşıp duruyor	. Tarihin (politik ek
lenmiş, hatta "mitosa" geri	dönüp duran	(Aydınlanmanın Di
ileri kadrolarına "çelişkiler"	üretip durmasından	başka bir anlam ta
mel eserler olarak "okura"	sunup dururken	, tahayyüllerimizdel
ınında tutunamayıp top gibi	yuvarlanıp durduklarından	çocukları gülmekt
nsiye gibi, bizimle koşarlar.	Yorulup durunca	onlar da gölgemizl
kuş türü başınızın üstünde	dolaşıp durur	: Beyaz akbaba, tu
hep iyiyi yaratan ruh ile	karşılaşıp duruyor	. Tarihin (politik ek
lenmiş, hatta "mitosa" geri	dönüp duran	(Aydınlanmanın Di
ileri kadrolarına "çelişkiler"	üretip durmasından	başka bir anlam ta
mel eserler olarak "okura"	sunup dururken	, tahayyüllerimizdel
e olmasa da 6 ay boyunca	dönüp durabilecek	. Mühendisler aync
ılır. Düşünün bir, dünyanın	dönüp durması	, gezegen ve yıldız
den fazla insan-öncesi türü	dolanıp duruyordu	. İşte insanlar şimd
övdesiye tramvayın önünü	kesip durmasını	sağlayacaktır. Ada
ki herkes her çeşit konuyu	tartışıp dururken	, kendi kendine bir
ınsına devam etti. Biz hâlâ	tartışıp duruyorduk	, "Yok efendim IM

8 Conclusions

This study demonstrates the application of NooJ dictionaries and cascaded graphs to Turkish as a highly-agglutinative language.
Inflection of multi-word units and proper names, syntactic grammars for disambiguation, statistical reports on the affixation of Turkish, a web interface to make further use of the module, additional constraints on Turkish derivation, optimization of the morphological graph are all plans for future releases.
For updates and related information please refer to companion website www.tudd.org

Keywords: Turkish, Nooj, corpus linguistics, POS tagging, grammatical tagging, morphological analysis, morphotactics, TNC

Acknowledgments
We thankfully acknowledge that this research was supported by a grant from TÜBİTAK (No.108K242).
The authors are also grateful to colleagues and students at Linguistics Department of Mersin University for their contributions to the module.

References

Akın, M. D., & Akın, A. A. 2007. Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi: ZEMBEREK. *Elektrik Mühendisliği, 431*, 38.

Bayraktar, Ö.& Taşkaya-Temizel, T. 2008. *Person name extraction from Turkish Financial news text using local grammar based approach.* In Proceedings of the International Symposium on Computer and Information Sciences.

Bisazza, A. 2009. *Designing a Nooj module for Turkish*. Paper presented at the Nooj Conference 2009.

Çiçekli, İ., & Temizsoy, M. 1997. Automatic creation of a morphological processor in logic programming environment. In *Proceedings of the 5th International Conference on the Practical Application of Prolog (PAP'97)*. London, UK.

Göksel, A., & Kerslake, C. 2005. *Turkish: A comprehensive grammar*. London & New York: Routledge.

Güngör, T. 2003. *Lexical and morphological statistics for Turkish*. Paper presented at the International Twelfth Turkish Symposium on Artificial Intelligence and Neural Networks.

Hankamer, J. 1989. Morphological parsing and the lexicon. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 392-408): MIT Press.

Kornfilt, J. 1997. *Turkish*. London; New York: Routledge.

Külekçi, M. O., & Özkan, M. 2001. Turkish word segmentation using morphological analyzer. In *Proceedings of EuroSpeech*. Aalborg, Denmark.

Lewis, G. L. 1967. *Turkish grammar*. Oxford: Oxford University Press.

Makedonski, P. 2005. *Finite state morphology: the Turkish nominal paradigm.* Universitat Tübingen, Tübingen.

Oflazer, K., Göçmen, E., & Bozşahin, C. 1994a. *An outline of Turkish morphology*: Technical Report, Middle East Technical University.

Oflazer, K. 1994b. Two-level description of Turkish morphology. *Literary and linguistic computing, 9*(2), 137-148.

Sebüktekin, H. I. 1974. Morphotactics of Turkish verb suffixation. *Boğaziçi Üniversitesi Dergisi, 2*, 87-116.

Sezer, E. 2001. Finite inflection in Turkish. In E. E. Taylan (Ed.), *The verb in Turkish* (pp. 1-47). Amsterdam: John Benjamins Publishing.

Underhill, R. 1976. *Turkish grammar*. Cambridge, Mass.: MIT Press.

Uzun, E., Uzun, L., Aksan, M., & Aksan, Y. 1992. *Türkiye Türkçesinin türetim ekleri: Bir döküm denemesi [Derivational suffixes of Turkish: A morpheme inventory]*. Ankara: Şirin.

Uzun, E. 1993. *Türkiye Türkçesinde sözlüksel yapı: Bir eleştirel çözümleme.* Ankara Üniversitesi, Ankara.

Uzun, E. 2008. Türetim eklerinin türetkenliğini ölçme önerileri üzerine. In Y. Çotuksöken & N. Yalçın (Eds.), *XX. Dilbilim kurultayı bildirileri 12-13 Mayıs 2006*. İstanbul: Maltepe Üniversitesi

Arabic Compound Nouns processing: inflection and tokenization

Ines Boujelben [(1)], Slim Mesfar [(2)], Abdelmajid Ben Hamadou [(3)]

[(1)]*MIRACL, University of Sfax, Tunisia*, Boujelben_Ines@yahoo.fr,
[(2)] *RIADI, University of Manouba, Tunisia*, mesfarslim@yahoo.fr,
[(3)]*MIRACL, University of Sfax, Tunisia*, abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

The abundance of Arabic compound nouns in medical corpora requires their listing in electronic dictionaries. However, the generation of all potential inflected forms as well as the recognition of agglutinated forms attached to each entry needs a special tokenization and inflection process due to the linguistic specificities these lexical entries. This paper describes a new approach for Arabic Compound Nouns inflection and tokenization processing. We use the linguistic platform NooJ and we build a set of linguistic resources related to the biomedical domain such as specialized dictionaries and automatic recognition rules. The experimental results are encouraging and highlight the proposed method; this method allows lexical coverage enhancement in our biomedical terminology recognition tool.

1 Introduction

The Arabic language and especially the biomedical written texts are rich with complex words. The construction of an Arabic Compound Nouns ACN dictionary can reduce the complexity of the text understanding, facilitate the automatic translation to other languages, etc. On the other hand, the ACN processing can remove the ambiguity of terms that facilitate the linguistic analysis. Finally, we can't neglect the importance of ACNtagging and the large coverage of lexical resources. This paper deals with three main parts. The first part introduces a new typology of Arabic compound nouns related to the biomedical domain. The second one is the identification of the problems related to the ACN and presents a study of the existing methodologies to process the compound words. Next, we explain our proposed method that will be experimented on a biomedical training corpus. Finally, we present the obtained results.

2 Typology of ACN for the biomedical domain

The compound nouns formulation concept presents a complex process where the entire completed research tasks try to define it. A compound noun is a consecutive sequence of at least two simple forms and blocks of separators (Silberztein, 1993b). In fact, the ACN can be a combination of different forms: a noun, an adjective and/or a particle. In this section, we will be interested in the compound nouns belonging to the biomedical domains. Most of ACNs are composed of one or more nouns (N), adjectives (ADJ), adverbs (ADV) or simple named entities (PR). We manually extract a list of about 30 patterns of ACN compositions (Figure.1).

Inspired of the old works (Kabbebi and al, 2006), (Addahdeh, 2001) and (Ben Hammouda and Haddar, 2009) and with the analysis of the different compound nouns existing in our medical corpora, we identified a lot of grammatical categories of Arabic compound nouns belonging to biomedicine.

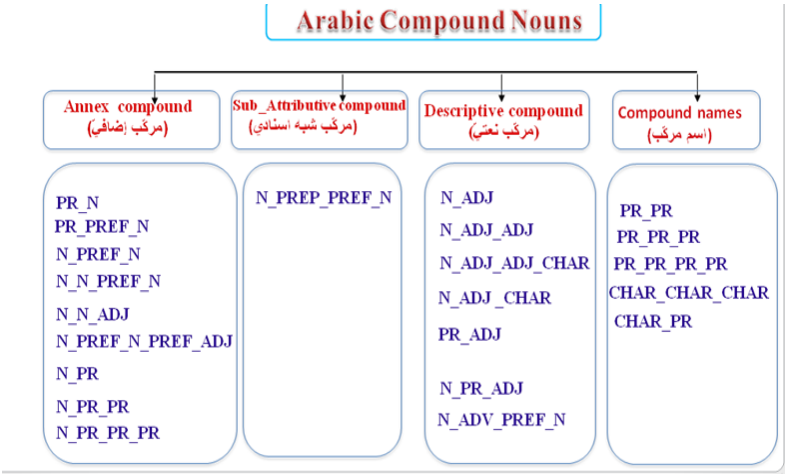


Figure 1. Typology of ACN

In fact, we recognized 4 main categories such as:

2.1 The Annex compound [إضافيمركب]

This syntagm is composed of indefinite noun (N) followed by a definite noun (PREF_N) or indefinite noun or invariant noun (PR) followed by indefinite noun. Indeed, we can find a noun followed by definite article (or not) followed by noun (N_N, N_PREF_N). It can be also made up of a combination of nouns or one or more than one NEs (N_PR, N_PR_PR, N_PR_PR_PR) like [اختبار إليزا, Test of Eliza].
Finally, this category can be presented by undefined nouns followed by other defined or undefined nouns and adjective (N_N_ADJ, N_PREF_N_PREF_ADJ) like for the (N_N_ADJ) category the example of medical exam[جِذْرِيّرَحمِاسْتِئْصَال, Radical hysterectomy]

2.2 The sub_attributive compound [اسناديشبهمركب]

This category is composed of a noun (N) followed by a preposition (PREP) and definite noun (PREF_N). It is designated by (N_PREP_PREF_N) such as the psychological diseases [الذاتعلىانطواء, autizm].

2.3 The descriptive compound [النعتيالمركب]

The descriptive compound is composed of two parts: the descriptive and the described part. For the described part, it can be presented by one or more adjectives (N_ADJ, N_ADJ_ADJ, N_ADJ_ADJ_ADJ).We must note that the descriptive will beconjugated in genre, in number, in identification and in vowellization with the described part.
This category can also be made up of noun and a named entity and adjective (N_PR_ADJ) or named entity followed by an adjective (PR_ADJ) like the name of diseases [إديما رئوية , lung Idema] or finally, it can be composed of a noun followed by an adverb followed by a definite noun (N_ADV_PREF_N) like for example
[فوقالبنفسجيةالأشعة, ultraviolet light].

2.4 The compound proper name [مركباسم]

The compound name is formed by the combination of invariants nouns that can be a character (CHAR) or named entity (PR) (not existing in dictionary) such asthe name of bacteria which is composed of (PR_PR) [أَسْتَبفْلُوكُوكِس أَبِيدِيرَمِيس, Ostavlocox Obedermis]. Unlike the other categories, this syntagm doesn't admit derivational forms.
We are only interested in the biomedical domain, for that, ACN can also be formed using other combinations of words.
These categories of ACN can have various inflected forms that depend on the name of syntagm.

3 Problems of Arabic Compound Nouns ACN

To introduce our work, we can notice that the medical language is rich with complex words. As we know, the compound noun is a noun that is made up of 2 or more words and the omission of any word can dismiss the meaning. In addition to that, the components of ACN can be separated by various separators like {« _ », « - », «" », « / »}. Let us take the example of the drug [أَنْتِزُوكُوكِس _ كِبلِس ,antrikoukis_pikalis] that can be represented in different forms in medical corpora [أَنْتِزُوكُوكِس _ كِبلِس , كِبلِس / أَنْتِزُوكُوكِس– كِبلِس , أَنْتِزُوكُوكِس | كِبلِس].
Furthermore, the Arabic words are characterized by their complex structure. The Arabic language is characterized by its very complex morphology because it is highly inflectional. Moreover, as we know, Arabic is an agglutinative language. Indeed, prepositions [فـً, in], conjunctions [و, and], articles [ال , the] and pronouns can be affixed to nouns, adjectives, particles and verbs.
Besides, we were faced to the problem of generating all the potential inflectional and derivational forms as well as the recognition of agglutinative forms that can be attached to each component of ACN. The following figure (Figure 2) shows an example of derivation and tokenization of the entry [قلبيةأزمة,«azma kalbyya», cardiac attack].

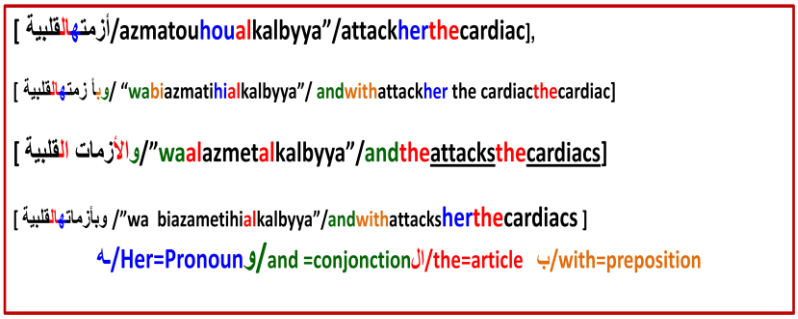


Figure 2. Derivation and tokenization

So, to solve these problems, there is a great need of a process that tokenizes each entry of ACNsdictionary.

4 Existing methodologies

Before explaining our proposed approach, we deal with the study of existing methodologies of tokenization and inflexion of compound words. In fact, we can distinguish two main potential solutions;

4.1 First approach

The first solution, focused on the assumption that both the compound and simple nouns are inflected and derived in a unified way (Silberztein, 2005). In fact, it is based on inflectional grammars. Here is an example of flectional forms of the derived nouns of [صدريقفص, « kafas sadri »] based on the writing of inflectional and derivational paradigms (Figure 3).

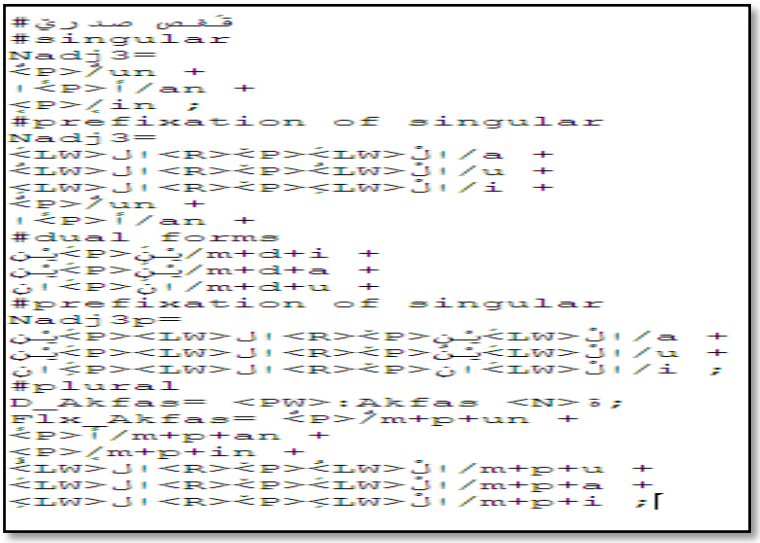


Figure 3. Inflectional and derivational paradigms

This first approach comes with an exponential complexity due to the various derived forms. As an example proving that, we can cite the treatment of the prefixation of compound noun. If we take the category of a descriptive compound composed of 2 words (N_ADJ), we will have 3 vowellization cases (genitive, accusative and nominative) in addition to 3 derived forms (singular, double and plural). For each derived and vowellization case, we can have four various ways of additional prefixations. As a result, we will have in total 36 paradigms (Table 1).

Number of components of ACN	inflectional paradigms			Total paradi-gms
	Vowled	Derived	Prefixed	
2 words (N_ADJ)	3	3	4 = 2	3*3*4=36
3 words (N_ADJ_ADJ)	3	3	$8=2^3$	3*3*8=72

Table 1. Total of inflectional paradigms

Consequently, to process all possible forms, we need a lot of manual works.So, what can happen to the complexity in the case of ACNs containing more components (3 or 4 components)?

4.2 Second approach

For the second approach, it focused on morphological grammar. Concerning this solution, we noticed that the use of the morphological NooJ grammars traditionally applied to tokenize the simple word forms was impossible because of the restriction of their use on only simple word forms.

Also, this method doesn't recognize the atomic unit that means it doesn't inherit the features of each component.

5 Proposed approach and implementation

In this paper, we propose a method allowing the tokenization of Arabic compound nouns through different steps as presentedin the following figure.

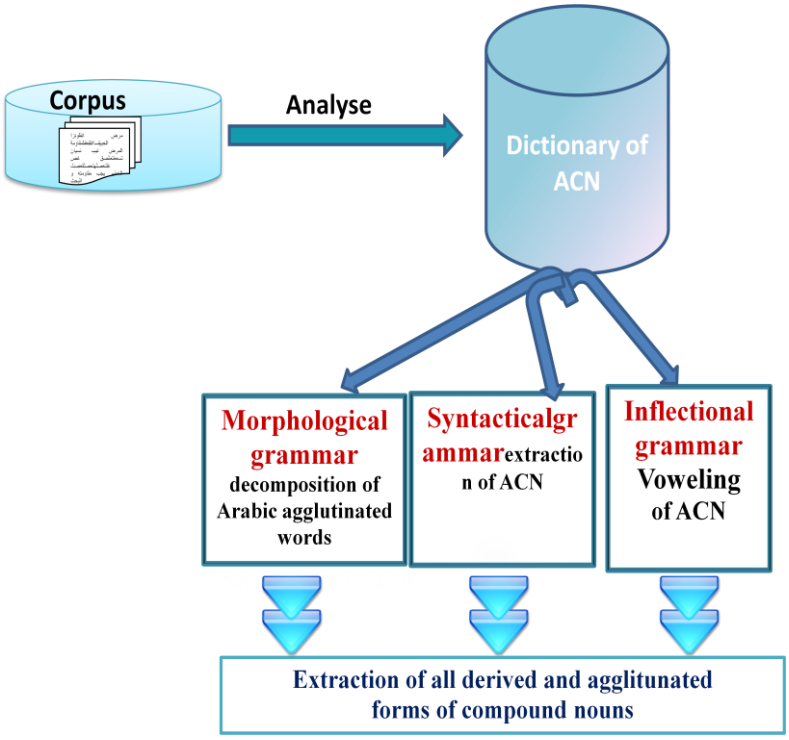


Figure 4. Proposed approach

Firstly, we start by building a corpus based on some texts belonging to the biomedical domain. Indeed, the fact of having a big corpus enables us to have the maximum of potential cases of Arabic compound nouns. After analyzing our corpus, we manually extract the compound nouns for building our dictionary of ACN. After that, we reach the study of the entries of dictionary in order to extract its compound grammatical category from which we deduce the typology of ACN [section 2]. Here is a print screen of our dictionary. As we see, we added semantic attribute referring to the category of ACN in order to treat each category separately in the syntactical grammar.

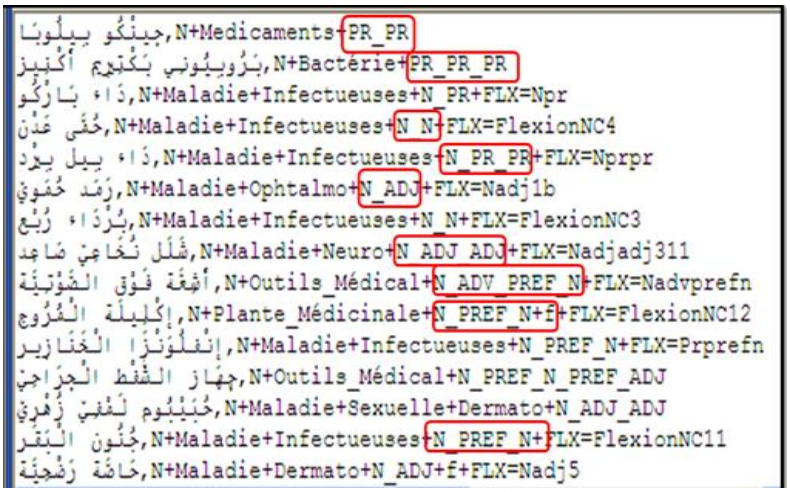


Figure 5. Building dictionary of ACN

In this step, we also study the different derivational, inflectional and agglutinative forms that can be attached to each entry.We need to elaborate 3 different types of grammars that will collaborate to give the expected results:

-Morphological grammar is used for decomposition of Arabic agglutinated words. We should note that we reused the grammar of agglutination (Mesfar, 2008).

- Inflectional grammar to generate the different voweled forms of the dictionary entries (genitive, accusative and nominative).

-Syntactical grammar to extract all related derived and agglutinated forms. The elaboration of syntactical grammar is illustrated by the following grammar that processes every category separately.

This grammar is composed of 17 sub-graphs where each one contains the appropriate treatments of the specific grammatical category.

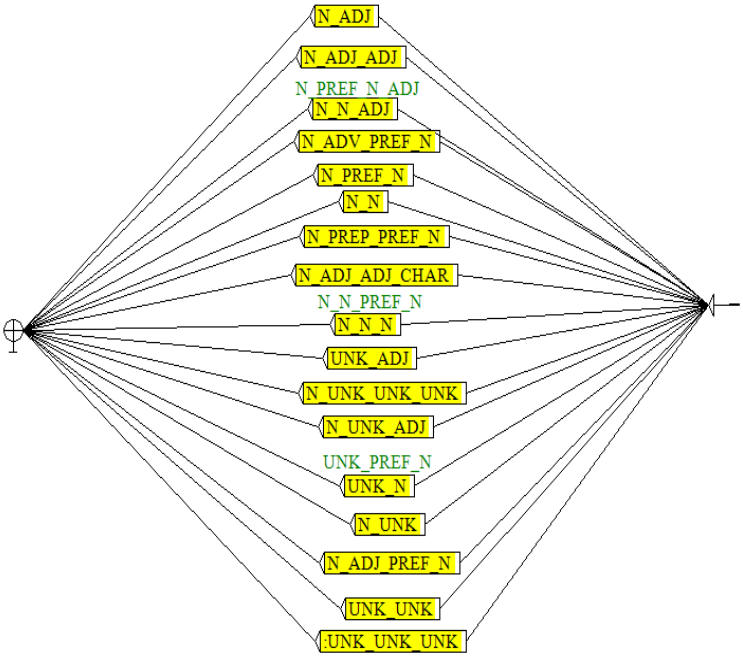


Figure 6. Local grammar of ACN processing

For further understanding, we will explain the embedded graph N_ADJ that treats the descriptive compound.
This local grammar recognizes all the derivational and agglutinative forms related to the descriptive category (N_ADJ).

In this embedded graph (Figure 7), we have three paths. We have stored the noun in the variable named ($Nom) and the adjective in the variable ($Adj). In the first path, we start with the embedded graph "Agglutination" which contains the list of prefixes that can be agglutinated to the nouns like prefix (definite articles[the/ال],prepositions [for/ل],conjunctions [and / و]. The node that contains the variable ($Nom) is linked with the node that contains a list of compound nouns" separators. This node is linked to the node that is stored in the variable ($adj). We add the constraint <$Nom_ $Adj_=:N+N_ADJ> to each path.

This constraint is validated by a dictionary lookup, in fact NooJ checks that the lemma of the noun (_$Nom) followed by space, followed by the lemma of the adjective (_$Adj) is listed as an ACN and has the category N_ADJ.

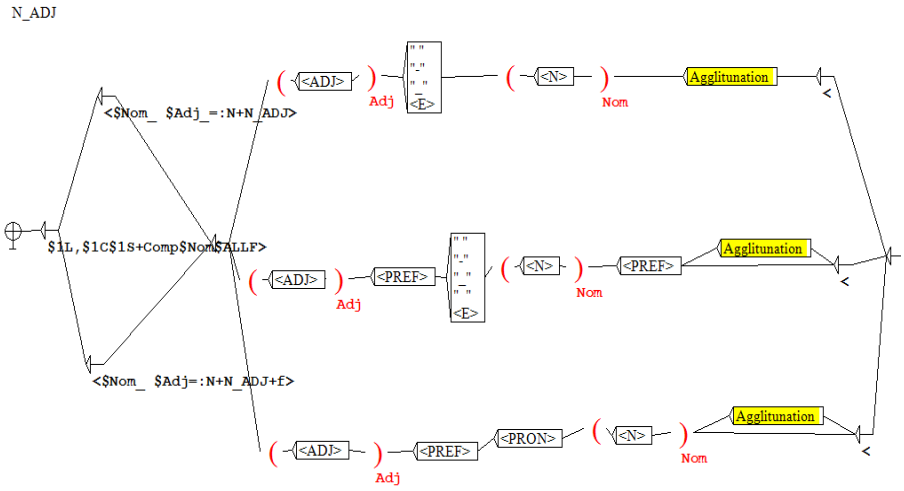


Figure 7. The sub grammar of the category N_ADJ

We must notice that we can sometimes find a descriptive compound that contains the descriptive part in the plural forms as a lemma in our ACN dictionary. For this reason, we added other constraints that are treated similarly and verified the existence of the lemma of nouns followed by the adjective stored in ($Adj) in our dictionary. If these constraints are checked, this grammar produces the resulting tag:

<$1L, $1C$1S+CompNomALLF>.

In this annotation, the variable $1L stores the lemma of the compound noun ($Nom_ $Adj). $1S stores the syntactic and semantic features, $1C stores the category of the compound noun and $ALLF can inherit all the inflectional properties of this noun.

To solve the problem of separators in compound nouns, in this grammar we added a node that contains the various separators between the words of compound noun.

6 Evaluation and results

To evaluate our system, we created a new corpus different from the training one that was used to manually extract rules. Here is an example of our Web corpus composed of 1003 articles (Figure 8), containing about 5159 text units and 107,140 different tokens.

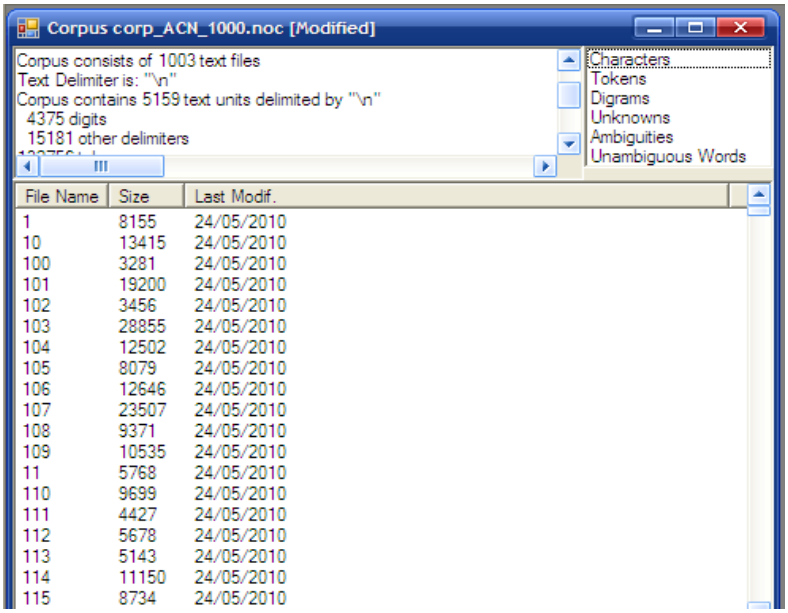


Figure 8. Building the test corpus

To evaluate the resulting module, we use the evaluation metrics like precision[1], recall[2]and F_measure[3]. Here is an example of the concordance obtained for the evaluation corpus.

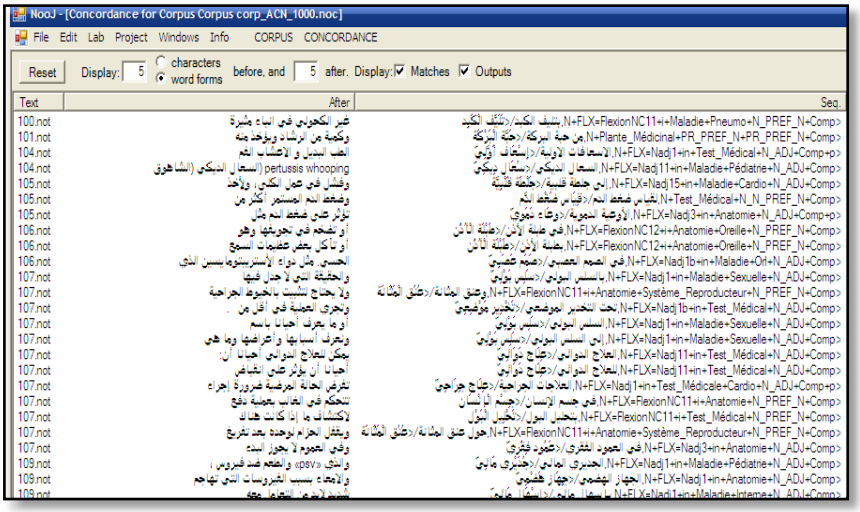


Figure 9. Concordance table obtained for the evaluation corpus

[1] measures the number of correct extractions

[2] defined as the number of correct extractions at or above P divided by the total number of correct extractions at all probabilities

[3] combines the measures of precision and recall into a single value.

Our system reached a precision of 98% and a recall of 90.5 % (Table 2).

Precision	Recall	F_mesure
98%	90,5%	94,2%

Table2. Results of evaluation

The performance results achieved were satisfactory in terms of precision, recall, and f-measure. In fact, the obtained results are encouraging and show that this process of inflection and tokenization of ACN can be used for compound nouns in various languages and domains.
The silence of our process of tokenization is due to different writing problems encountered in journalistic articles such as the double space that can separate words of compound nouns.

7 Conclusion and perspectives

In this article, we presented our process of tokenization and derivation applied to the Arabic compound nouns. We are interested in the biomedical domain. To achieve our approach, we have to elaborate an approach based on transducers, dictionaries and biomedical corpus made by means of NooJ. We studied the existing applications processing the ACNs and recognized their disadvantages. These resources are experimented on a corpus belonging to the biomedical domain different from the first studied corpus. The obtained results are encouraging with a rate of 98% for the precision and a recall of 90, 5%. As perspectives, we intend to orient our work to the extraction of Arabic compound nouns of any domain and any language.
Also, in reviewing the task of terminological extraction, we can notice that the named entities extracted appear in various categories of compound nouns. For that, we suggest replacing local grammar in some named entities by compound nouns dictionary using our process of tokenization.

References

Addahdah A., 1992, librairie Liban, Bairout, « قواعد اللغة العربية في جداول و أوقاتمعجم», 114-116.

Ben Hammouda F et Haddar K, “An Arabic dictionary for compound nouns “, NooJ ,Juin 2009.

Kabbebi S., A. Bennour, S. Missaoui et M. Chaouch, 2006. “*Manuel scolaire de langue de 8ème année de bas*e”, Ed 2006, 31-92.

Mesfar Slim, 2008. “ Analyse Morpho-syntaxique Automatique et Reconnaissance Des Entités Nommées En Arabe Standard”. Thesis”, "Graduate School" ―Languages, Space, Time, Societies", Paris, France.

Silberztein, Max. 2005. “NooJ's Dictionaries”. *In the Proceedings of the 2nd Language and Technology Conference*, Poznan.

Silberztein Max., 1993. “Les groupes nominaux productifs et les noms composés lexicalizes”. In Lingvisticae Investigationes XVII: 2, Amsterdam: John Benjamins B.V

Morphology based recognition of Greek verbs with Nooj

Angeliki Efthymiou[(1),] Zoe Gavriilidou[(2)]

[(1)(2)]*Democritus University of Thrace*

1 Introduction

The aim of this study is the automatic recognition of Modern Greek derived verbs, in order to ameliorate text annotation of the Greek Nooj module. In the first part, we will present some information about the productivity, the frequency and the semantic properties of these verbs. For the semantic analysis of the derived verbs we will use the associative morphological model of Corbin (1987, 1991) and the theory of lexical conceptual semantics of Jackendoff (1990). The combination possibilities and restrictions between semantic classes of bases and given suffixes which will follow will be based on the theory of Classes of objects(Gross 1992). Finally, a demonstration of derivational rules and productive morphological grammars constructed for the automatic recognition of Greek derived verbs will close this paper.

2 Modern Greek verb-forming morphemes: productivity-frequency

Modern Greek has seven main verb-forming suffixes and one main semi-suffix, namely the element *–ποιώ*:

–ίζω: *βουρτσίζω* 'to brush', *μαυρίζω*'blacken',*αριστοτελίζω* 'imitateAristotle'
-(ι)άζω: *ρυτιδιάζω* 'to wrinkle, become wizened'.
–όνω : βουτυρώνω 'to butter', *μαλακώνω*'soften',*καρφώνω*'to nail'
–εύω : προεδρεύω 'to chair, preside',*αγριεύω*'make/become fierce, get/make angry', *ταξιδεύω* 'to travel'
–αίνω : χοντραίνω 'get/grow fat, thicken' *λιπαίνω* 'lubricate, fertilize'
–άρω : στρεσάρω 'to stress',*πουδράρω* 'to powder',*ζουμάρω* 'to zoom'[1]
–ποιώ : γραμματικοποιώ 'grammaticalize',*απλοποιώ* 'simplify', *περιθωριοποιώ* 'marginalize'[2]
-ύνω: *δασύνω* 'aspirate'

Due to the absence of systematic investigations for the frequency and productivity of Modern Greek affixes, in this study we present two kinds of empirical data, namely
a) *–ίζω, –(ι)άζω, –όνω, –εύω, –αίνω, –άρω, -ύνω* and *–ποιώ* verbs which are listed in the *Reverse Dictionary of Modern Greek* (RDMG) (cf. Efthymiou (2010 and to appear)),

[1] The suffix – *άρω* is of Italian etymology. It is attached mainly to nominal bases of non-Greek origin (cf. Αναστασιάδη-Συμεωνίδη (1994), Efthymiou (to appear)).
[2] The confix *–ποιώ* developed from the Ancient Greek verb ποιῶ 'make/do' (cf. Αναστασιάδη-Συμεωνίδη (1986), Γιαννουλοπούλου (2000)).

b) *–ίζω, –(ι)άζω, –ώνω, –εύω, –αίνω* and *–άρω* verbs which are present in Printed School Modern Greek (as investigated in Efthymiou, Havou and Gavanozi (2010), cf. also Efthymiou 2010)[3].

Our data from RDMG and printed school Modern Greek are presented in tables (1) and (2):

verbs in	raw data	scrutinized data
–ίζω	3507	approx. 650
–(ι)άζω	2260	approx. 313
–ώνω	2106	approx. 500
–εύω	1207	approx. 325
–άρω	547	approx. 150
–ποιώ	252	approx. 200
–αίνω	687	approx. 100
-ύνω	186	approx. 32

Table 1.*Verb forming processes (type frequency): Data extracted from RDMG*

words in	Token frequency in printed school MG
–ικός (adj.)	12 %
–ση(n.)	11,1 %
–ία (n.)	9,6 %
–ίζω (v.)	9,5 %
–α (adv.)	8 %
–ώνω (v.)	5,6 %
–εύω (v.)	5,6 %
-μα (n.)	4,9 %
-άζω (v.)	3,9 %
-ιάζω (v.)	2,5 %

Table 2. *token frequency in printed school Modern Greek*

Although we do not claim that our data is thoroughly reliable or could be generalized to any kind of textual typology, the figures in tables(1)and (2) allow the following generalizations (cf. Efthymiou 2010 and to appear). Firstly, as expected according to the literature on productivity (cf. see Baayen 2008, Bauer 2001, Plag 1999), Modern Greek suffixes seem to differ considerably in their type and token frequency. Secondly, of all verb-forming suffixes *–ίζω* seems to be the most productive. Thirdly *–ύνω* and *-αίνω*do not seem to be synchronically productive in Modern Greek.

[3] In Efthymiou, Havou and Gavanozi (2010), 54 Modern Greek suffixes were investigated. The material collected from the corpus of the Textbooks of language and literature, history, mathematics, religion, and environmental education of the 3rd grade Primary School contains 7773 tokens of Modern Greek suffixed words. Notice, however, that *–ύνω* and *–ποιώ* were not included in this study. Moreover, *–άζω* and *–ιάζω* were analyzed as variants of the same suffix.

3 The semantics of verb forming processes

As shown in Efthymiou (2010andto appear), these derivatives show a wide variety of meanings, such as causative, resultative, inchoative, ornative, locative, instrumental, performative, similative, etc.In particular the meanings of these Modern Greek verb-forming processes can be summarized in table (3) (cf. Efthymiou 2010, to appear):

	–ί ζ ω	–(ι) ά ζ ω	– ώ ν ω	– ε ύ ω	– α ί ν ω	– ά ρ ω	– π ο ι ώ
cause to become x	✓	✓	✓	✓	✓	✓	✓
become x/be provided with x	✓	✓	✓	✓	✓	✓	
make x go to/in/on something	✓	✓	✓	✓	✓	✓	✓
make something go to/in/on x	✓	✓	✓	✓		✓	✓
do x	✓	✓ (?)		✓		✓	
do /act like x	✓	✓ (?)		✓		✓ (?)	
use x	✓	✓ (?)	✓	✓		✓	
carry out the official activities of x				✓		(?)	

Table 3. *The meanings of –ίζω, –(ι)άζω, –ώνω, –εύω, –αίνω, –άρω derivatives and –ποιώ formations*

As shown in Efthymiou (2010 and to appear), although Modern Greek verb-forming suffixes seem to share the same underlying structure, each suffix seems to develop a semantic category prototype related to the frequency of the meanings expressed by the derivatives. Thus, not all semantic categories are equally possible or probable for all verb forming processes. The suffix *–ίζω* is more probable to participate in similative, performative and instrumental (or manner of motion) interpretations[4], whereas the suffix *–ιάζω*[5], is more probable to express an inchoative meaning. Furthermore, the suffix *–ώνω* is more likely to convey ornative or instrumental and causative meanings, while the suffix *–εύω* is more likely in similative/stative/essive and inchoative interpretations. Finally, *–ποιώ* is more likely to have a resultative meaning. Furthermore, *–ιάζω* seems to be the prevailing default verb forming suffix for the inchoative interpretation 'become/become provided with unwanted x', whereas the suffix *–εύω* seems to be prototypically associated with the similative/stative/essive meaning 'carry out the official activities of x'. Finally, *–ώνω* seems to be prototypically associated with the ornative meaning, whereas *–ίζω* seems to be the default verb forming suffix for the similative meaning 'act like' (cf. Efthymiou to appear and 2010).

Moreover, Modern Greek suffixes don't seem to select the same type of base. For example, –ίζω is basically the only suffix among the suffixes of our study that attaches to onomatopoetic words, proper names and names of colours. On the other hand, *–εύω* is the only suffix that attaches to stage-level nouns denoting offices of persons, for deriving verbs with the meaning 'carry out the official activities of x for a certain period' (cf. Efthymiou (to appear and 2010)). Finally, *–(ι)άζω* is the only suffix that attaches to numerals and -*ποιώ* is basically the only verb forming morpheme in our study that attaches to relational adjectives in *–ικός*.

[4] For the suffix *–ίζω*, see also Charitonidis (2005).

[5] For the suffix *–(ι)άζω*, see also Efthymiou (2010)

4 Local Grammars for the automatic recognition of the derived verbs in Greek

As mentioned above, some of the suffixes studied in this paper remain very productive in Modern Greek and create a high number of neologisms which, when found in texts, cause problems in text annotation. To avoid overloading the dictionary by introducing all the above mentioned examples as separate entries, Nooj's annotations can represent suffixes described by derivational rules or productive morphological grammars which describe the formation and facilitate recognition of such examples (Silberztein 2003).
For example, the following grammar (see figure 1) was created for the automatic recognition of the present tense of verbs in *-ίζω, -(ι)άζω, -εύω, -ώνω, -άρω, -αίνω, -ύνω,* and the confix *–ποιώ*.

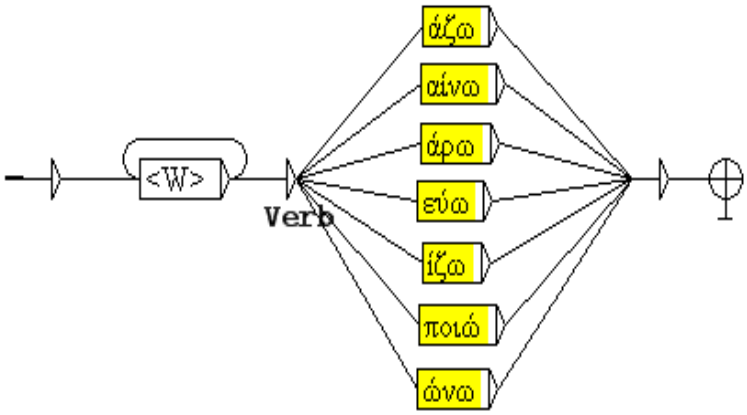


Figure 1: *Verb recognition*

Greek is a strongly inflective language disposing for each verb an extremely high number of different forms of inflexion. Consequently, after the recognition of the simple present of verbs in *ίζω, -(ι)άζω, -εύω, -ώνω, -άρω, -αίνω, -ύνω,* and the confix *–ποιώ*it was important to recognize these verbs in all the other tenses and modes as well as in all inflected forms of present tense. To achieve the automatic recognition of inflected forms, eight different grammars, one for each verbal suffix, were created as illustrated in figures 2 and 3.

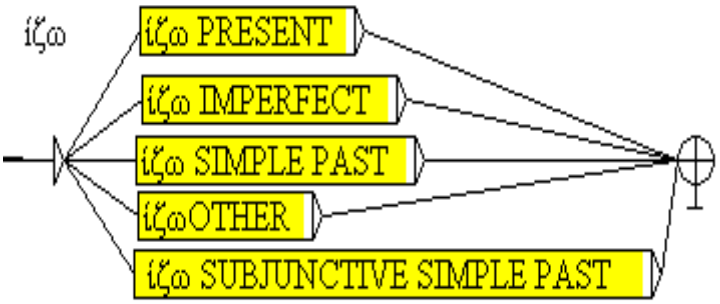


Figure 2. *Tense recognition of the verbs in –ίζω*

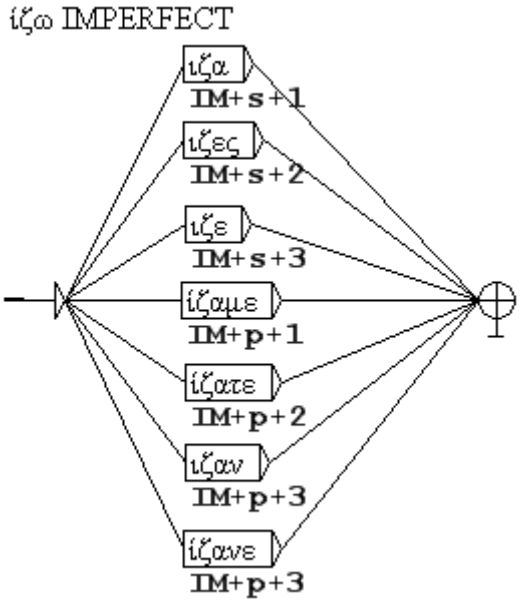


Figure 3.*Tense recognition of the verbs in –ίζω*

After the application of the tense recognition grammars, problems of ambiguity arose between the first singular imperfect person of verbs in *–ίζω*and *–άρω* and the singular nominative, accusative and vocative of nouns ending in *ιζα* and *αρα*[6] (e.g. σκού**πιζα** 'I was shoveling'-**πίπιζα** 'flute') as well as between the second singular imperfect person of verbs in *–άρω*or non analyzable loan verbs (e.g. μαρκάρω 'το mark') and the plural nominative, accusative and vocative of nouns ending in *αρα* (σούτ**αρες**'you were shooting – κάμ**αρες**'chambers'). As the number of nouns ending in *ιζα* and *αρα* was restraint, we chose to resolve this ambiguity by adding the list of these nouns as counter-examples in the grammars of tense recognition of the above mentioned verbs.

[6] In *πίπιζα* and *κάμαρα* the suffixes *–ίζω* or *-άρω* do not feature: *ιζα* and *αρα* are sequences of letters including the inflection suffix *–α*.

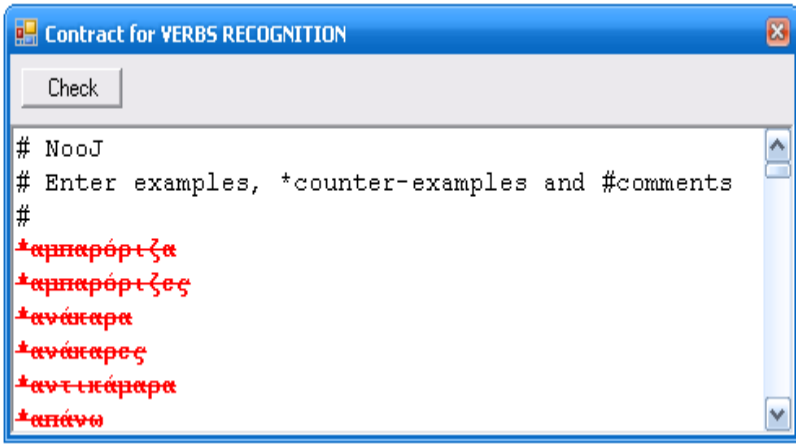


Figure 4. *Ambiguity resolution of the imperfect tense forms of verbs in –ίζω and -άρω*

Homonyny (more precisely homography) and in some cases ambiguity problems arose also for the second singular future person of verbs ending in *–άζω*, *-ίζω*, *-ποιώ* and*–ώνω* and the plural nominative, accusative and vocative of nouns in *–ση*, *-ποίηση* and *–ωση* (δηλητηρι**άσεις** 'you will poison/poisonings, σκουπ**ίσεις**'you will shovel'–κλ**ίσεις** 'inflections', πραγματοποι**ήσεις** 'you will accomplish/accomplishments, ματ**ώσεις** 'you will bleed'-κακ**ώσεις** 'lesions') which are numerous in Greek language. For the resolution of this type of homonymy and ambiguity, it was chosen to create a disambiguation grammar taking in consideration the syntactic context where a noun or a verb could appear (see Figure 4).

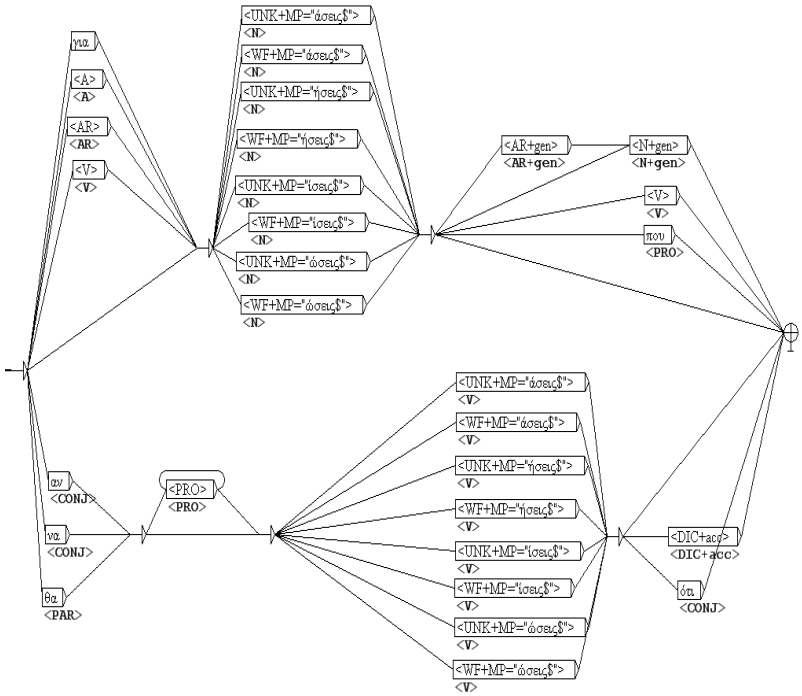


Figure 5. *Homonymy and ambiguity resolution of the future tense forms in –άζω, -ίζω, -ποιώ and–όνω*

After the application of the above mentioned disambiguation grammar all cases of homonymy and ambiguity not only for verbs existing in the macrostructure of our dictionary but also for neologisms or derived verbs not found in the dictionary were resolved as illustrated in Figures 6 and 7. Moreover this disambiguation grammar helped recognize all the misspelled verbs found in texts.

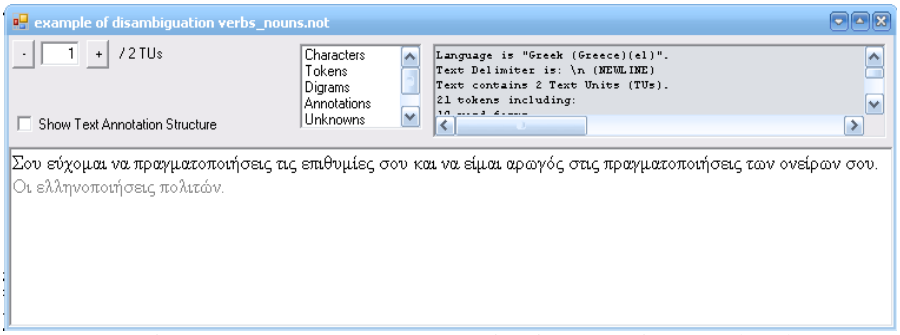


Figure 6. *Example ofHomonymy and ambiguity resolution of the future tense forms in –άζω, -ίζω, -ποιώ and–όνω*

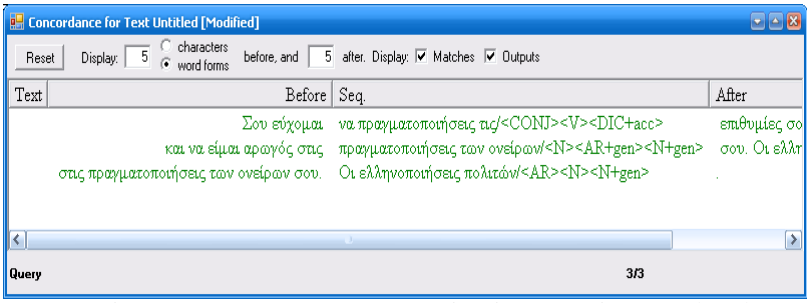


Figure 7. *Example ofHomonymy and ambiguity resolution of the future tense forms in –άζω, -ίζω, -ποιώ and–ώνω*

The aim of the present study was also to give to Greek Nooj module the possibility to make a morphological analysis of the derived verbs in *-ίζω, -(ι)άζω, -εύω, -ώνω, -άρω, -αίνω, -ύνω*, and the confix *–ποιώ*. For that reason, a number of morphological grammars were created for analyzing the derived verbs in stems and suffixes or confixes (see Figure 8).

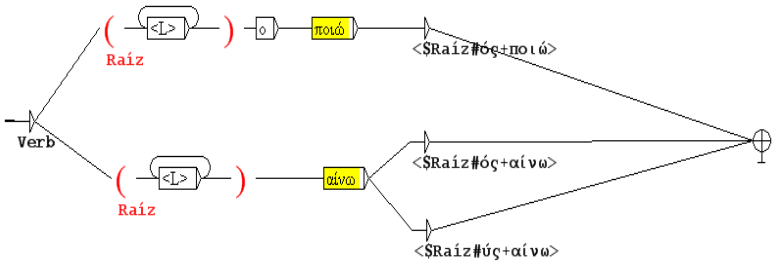


Figure 8. *Morphological analysis of verbs in -ποιώand –αίνω*

Cases of variants of stems were also taken under consideration as exemplified in Figure 9.

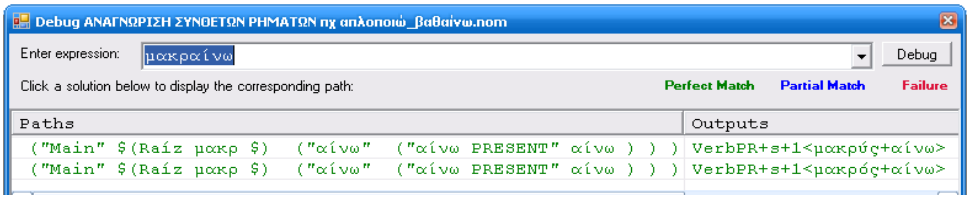


Figure 9. *An example of morphological analysis of verbs in –αίνω*

This possibility can be further exploited in L1 or L2 teaching for the elaboration of pedagogical applications in order to teach word formation in Greek.

Finally, we attempted meaning assignment to annotated words byconstructing grammars like the one presented in Figure 10. The retrieval of semantic information was not possible for all the examples encountered because of irregularities in meaning construction in Greek derived words. More precisely, there was not a one to one relation between stems, suffixes and meanings.

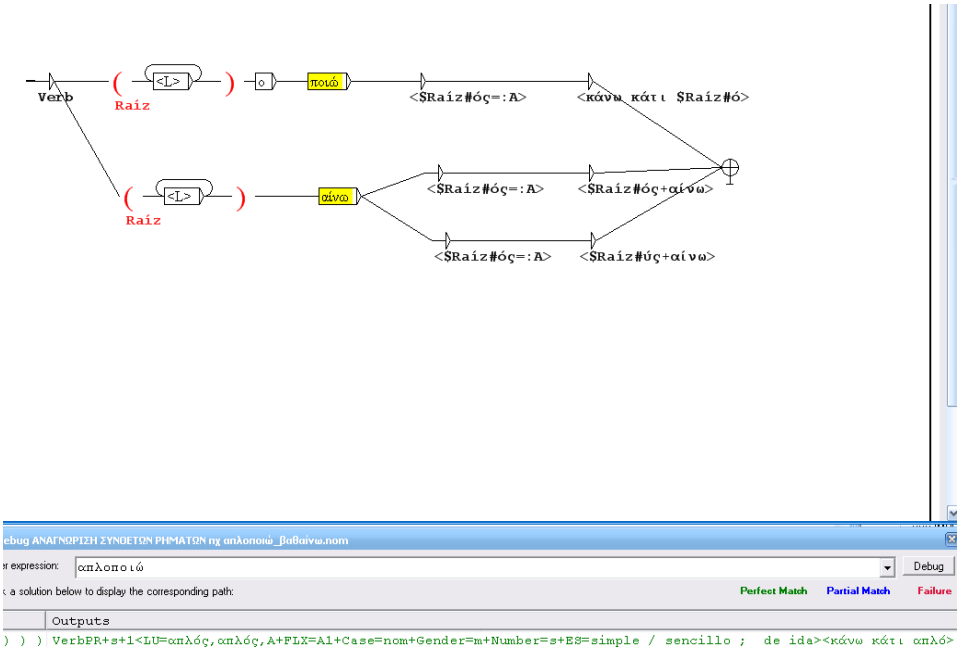


Figure 10. *Semantic analysis of verbs in –ποιώ*

5 Conclusions

The purpose of the present paper was the automatic recognition of Modern Greek derived verbs using the Nooj annotation tools and grammars. We constructed a number of Nooj grammars for the recognition, homography resolution, morphological analysis and semantic analysis of Greek derived verbs which helped us avoid overloading our dictionary. Future work would insist on the semantic analysis of Greek derived verbs by formalizing rules predicting the combination possibilities and restrictions between semantic classes of bases and given suffixes.

References

Anastassiadis-Symeonidis A. 1986. "Η φύση και η παραγωγικότητα του σχηματιστικού στοιχείου -ποιώ". *Studies in Greek Linguistics* 1985: 49-70.

Anastassiadis-Symeonidis A. 1994. *Νεολογικός Δανεισμός της Νεοελληνικής: Άμεσα Δάνεια από τη Γαλλική και την Αγγλοαμερικανική*. Thessaloniki.

Anastassiadis-Symeonidis A. 2002. *Αντίστροφο Λεξικό της Νέας Ελληνικής*. Thessaloniki: Institouto Neoellinikon Spoudon.

Baayen H. 2008. 'Corpus linguistics in morphology: morphological productivity.' In Ludeling A. and Kyto M. (eds.) *Corpus Linguistics. An international handbook*. Mouton De Gruyter: Berlin, 899-919.

Bauer L. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Corbin D. (1987). *Morphologie dérivationelle et structuration du lexique*. 2 Vols. Tübingen: Niemeyer.

Efthymiou A. 2010. "How many factors influence the meaning of denominal verbs? The case of Modern Greek verbs in –(i)ázo." Talk presented at the Workshop 'Meaning and Lexicalization', 14th International Morphology Meeting, Budapest 13-16 May 2010.

Efthymiou A. (to appear) "The Semantics of Verb forming Suffixes in Modern Greek". In the *Proceedings of the 19th International Symposium of Theoretical and Applied Linguistics*, 3-5 April 2009, School of English, Aristotle University of Thessaloniki.

Efthymiou A., Gavanozi V. & Havou E. 2010. 'The frequency of Modern Greek suffixes in the primary school textbooks'. Talk presented to the conference *Ζητήματα διδακτικής της γλώσσας* [Issues of Language Teaching] Democritus University of Thrace, Komotini, 7-8 May 2010.

Gaeta L. & D. Ricca. 2003. 'Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data'. *Rivista di Linguistica* 15/1: 63-98.

Gianoulopoulou G. 2000. *Μορφοσημασιολογική σύγκριση παραθημάτων και συμφυμάτων στα Νέα Ελληνικά και τα Ιταλικά*. PhD Dissertation. University of Thessaloniki.

Gottfurcht C. 2008. *Denominal Verb formation in English*. PhD Dissertation, Northwestern University, Evanston, Illinois.

Gross G. 1992.Forme d'un dictionnaire électronique. *Actes du colloque La station de traduction de l'an 2000*. Mons.

Jackendoff R. 1991. *Semantic Structures*. Cambridge: MIT Press.

Lieber R. 2004. *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Plag I. 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin/N. York: Mouton de Gruyter.

Plag I., Dalton-Puffer, C. & Baayen H. 1999. 'Productivity and register.' *English Language and Linguistics* 3: 209-228.

Silberztein M. 2003. *NooJ Manual*, available at the web site http://nooj4nlp.net (200 pages).

NooJ disambiguation local grammars for Arabic broken plurals

Samira Ellouze[(1)], Kais Haddar[(2)]& Abdelhamid Abdelwahed[(3)]

[(1)(2)]MIRACL, university of Sfax, Sfax, Tunisia
[(3)]LSCA, university of Sfax, Sfax, Tunisia

Abstract

The ambiguity in the detection of broken plural is a part of morph-syntactic ambiguity. We present here a linguistic approach for the elimination of ambiguities between Arabic broken plural and the other grammatical categories. To apply this approach, we use morphologic and syntactic grammars. These grammars are presented by the finite state transducers of the NooJ platform. With this platform, we experiment and evaluate the proposed approach.

1 Introduction

One of the problems encountered by researchers in the domain of NLP is without doubt the morphological ambiguities, meeting with the construction of a tagger. It is widely spread in all natural languages; a word may have several distinct meanings. Human can choice the suitable meaning according to the context in which the word occurs but computer cannot choice it without adding specific treatments to morphological analysis. The problem of morphological ambiguity doesn't stop in morphological analysis; it affects all other analysis (e.g., syntactic analysis, semantic analysis).

The detection of broken plural for simple nouns is a part of the morphological analysis. But we do not always correctly detect this type of plural. Sometimes, we find ourselves in a situation of conflict or ambiguity. This conflict is caused by different morph-syntactic phenomena related to the Arabic language as the lack of vowels, agglutination, etc.

In this context, we propose a linguistic approach to reduce or eliminate the ambiguities between broken plural and other categories. In order to experiment and evaluate this approach, we use the linguistic platform NooJ.

In this paper, we present first different approaches of morph-syntactic disambiguation. Then, we enumerate ambiguity causes. After, we implement the proposed method of disambiguation using the linguistic platform NooJ. Finally, we perform an experiment and an evaluation of the achieved tool.

2 Background and previous works

There are many works that are dealt with the problem of morphological disambiguation. Among these works, we cite Laporte (1999), Freeman (2001), Tlili (2006) and Amardeilh (2007). Each work of them uses either the linguistic approach or the statistical approach or both at once (Hybrid approach).

The linguistic approach of disambiguation is based on system of rules. The rules are generally divided into four types: grammatical, structural, semantic and logical. The grammatical constraints are most used to lift the ambiguity. All types of constraints mentioned above are divided into three categories: contextual rules, heuristics and non-contextual rules. Among Arabic taggers which are based on the linguistic approach, we cite the tagger of Freeman (2001). Also, we find other taggers that apply to other languages such as Brill tagger in Brill (1992, 152-155), VISL tagger in Karlsson et al. (1995) and ELAG tagger in Laporte (1999).

The development of different rules to disambiguate requires much time and effort. This has prompted researchers to use statistic approach.In this approach most of time, lifting ambiguity is done using for example the hidden Markov models according to Merialdo (1994). Various taggers have used statistic approach to implement disambiguation process. Among the Arabic taggers based on this approach, we cite the Al Shamsi tagger in Al Shamsi (2006, 31-41). Also, there are a larger number of taggers for other languages for example TnT tagger presented in Thorsten (2000) for English and German. And TreeTagger presented in Schmid (1994) designed for English. Note the last tagger is based on the binary decision tree.

The given results show the shortcomings of statistic approach, which has prompted researchers to adopt a hybrid approach that combines linguistic approach with statistic approach. This approach benefits from the advantages of the two previous approaches. The majority of disambiguation modules implemented in commercial systems combines several techniques to increase the performance of their taggers. Among Arabic taggers which used hybrid approach, we cite the tagger proposed in Elhadj (2009). This tool has been proposed to tag Koran. Also, we cite APT tagger in Khoja et al. (2003). The taggers applied on other languages, we mention MBT tagger in Walter (2002) which is designed for English language. We also find XeLDA tagger in Amardeilh (2007) designed for English and other languages.

3 Broken plural and causes of ambiguities

The broken plural named also internal does not obey to specific rules. According to Abbes (1984)and Ellouze et al. (2009), the internal structure of a singular word is changed to obtain a plural form. This change is either by adding one or more letters, or by removing one or more letters or by changing the vocalization or whether the combination of several of these cases. For example, the singular word "مصباح" (*mis□bāh□*, lamp) has the broken plural "مصابيح" (*mas□ābīh*, lamps). The broken plural has many forms, more than 40, usually unpredictable. These forms are divided into two categories: plural of paucity which is used in cases where there are more than three and less than ten items and plural of multiplicity which is used in cases where there are more than ten items.

It is true that the specificity of broken plural makes its detection difficult and ambiguous. But there are other causes that increase the ambiguity of the detection of this plural. In many cases, we find ourselves in a situation of conflict or ambiguity. This conflict is caused by different morpho-syntactic phenomena related to the Arabic language as:

- **Agglutination:** Arabic language can have agglutinative forms which contain: articles, conjunctions and prepositions at the beginning of the word and pronouns at the ending of the word. In some cases, these agglutinative forms do not allow the detection of

broken plural correctly; this is due to the identical form of some agglutinative forms with not agglutinative forms in the case of lack of short vowels. For example, the unvowelled word "بقول" (buqūl), it can be treated as a singular noun in agglutinative form where "ب" (bi - with) is a proposition and "قول" (qawl, saying) is the singular noun; or a broken plural noun (buqūl, legumes).

- **Lack of short vowels:** The lack of short vowels in the Arabic texts is among the main causes of the ambiguity faced in the detection of broken plural. Most Arabic texts (books, newspapers, magazines ...) are not vowelized or partial vowelized. Only the Quran and the children books are totally vowelized. With the absence of short vowels word can be vowelized with different ways and can take different grammatical forms. This causes different types of ambiguities (e.g., lexical ambiguities, morphological ambiguities).

As we have seen previously, agglutinative forms are a source of ambiguity but it can be also a source of disambiguation. In fact, it is a source of disambiguation because some enclitic or proclitics can only be attached to noun or adjective. In the following section, we will see how to use agglutinative forms in disambiguation.

4 Proposed method

In this work, we propose a method permitting the disambiguation of the Arabic broken plural according to Ellouze (2010). This method is based on a linguistic approach of disambiguation. In this approach, there are three types of disambiguation rules: context rules, heuristics, not context rules. To apply this approach, we identify grammatical rules that distinguish between a broken plural word and other possible cases. These rules are taken from Arabic grammar books such as Addahdah (1992) and Niama (2000). The identified rules are applied either in syntactic analysis or in morphological analysis. The application of some rules in morphological analysis is due to agglutinative forms. For this raison, we compose this method to two steps: morphological disambiguation and syntactic disambiguation. The follow figure represents these two steps.

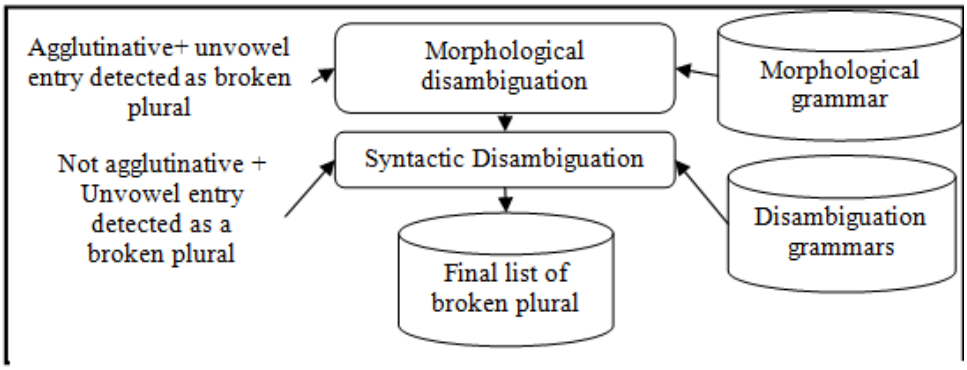


Figure 1. Proposed method steps

The figure 1 shows the two steps are based on two types of grammars: morphological grammar and syntactic grammar.

Morphological disambiguation

In this step, we only eliminate ambiguities of agglutinative word. The ambiguity of declination can be totally or partially removed in this step. This is depending on the composition of agglutinative word. The unvowel word initially known as broken plural will be "معربة" (*muʿarrabat* - declines) with three possible cases of declination: "مجرور" (*mağruwr* - genitive), "منصوب" (*mansuwb* - accusative) or "مرفوع (*marfouʿ* - nominative). But these declinations will decrease when the word is recognized as part of an agglutinated form. The decrease of declination depends on prefixes (preposition, conjunction…) and suffixes (pronoun) which they've been glued to the initial word recognized as broken plural. This is done by adding appropriate restrictions to the word according to agglutinated form of initial entry. For example, the following figure shows the declination of the word "بيوت" which is a part of the agglutinative word "كبيوتهم" (*kabuyūtihim*, like their houses).

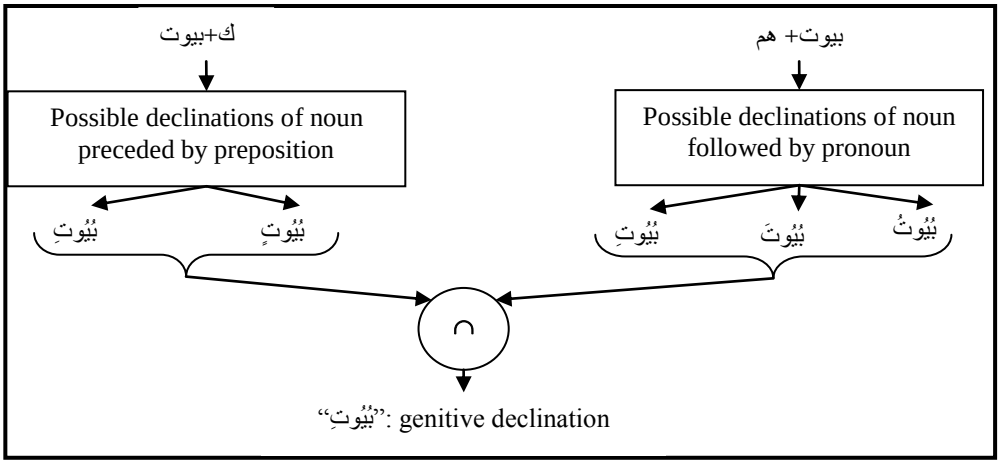


Figure 2. Example of output of an agglutinated and unvowelized input

In figure 2, we show that with the presence of the prefix "ك" (*ka*, as), we only retain declination of "مجرور" (*mağruwr* - genitive) case. And with the presence of the personal pronoun "هم" (*hum*, their), we remove the tanwin ً -(in) that is used in the case of an undefined noun since the personal pronoun makes the word defined. In the end, we find as a result only the word بيوتِ (*buyūti*).

Syntactic disambiguation

In the step of syntactic disambiguation, we use syntactic grammars of disambiguation for removing certain ambiguities encountered due to non-vowels texts on the one hand, and the presence of agglutinated forms, on the other hand. Indeed, we find many Arabic words that become the same when we remove the vowels. As we have seen previously, there are three

types of disambiguation rules: context rules, heuristics, not context rules. Thereafter, we give some examples of syntactic disambiguation grammars.

- The context rules: these are rules of restriction which were extracted from the books of Arabic grammar such as Addahdah (1992) and Niama (2000). Here are examples of these grammars:

  We present below our first grammar named GL_1. This grammar allows for the disambiguation of prepositional phrase. This phrase consists of a preposition followed by a noun followed by a pronoun.

G_{L1}	<phrase>	→	<preposition> + <word>+ <pronoun>
	<preposition>	→	"تحت", "فوق", "خلف", "إلى",…
	<word>	→	<noun>| <adjective>
	<adjective>	→	"أصدقاء",…
	<noun>	→	"شواطئ", "منازل", "كتب", "أوراق",…
	<pronoun>	→	"ه", "ي", "كم",…

Let the following example to which we apply the GL_1 grammar. This grammar is used to remove the ambiguity found in the prepositional group of Example 1.

Example 1:

وجدت نظارة أختي تحت كتبي

wajadtu naz□ārata ʾuhtiy tah□ta kutubiy

I found my sister's glasses under my books

The word "كتب" in the sentence of Example 1, can have two different categories: noun or verb. Applying grammar GL_1, the category verb will be rejected since after a preposition (in this case "تحت" (*tah□ta*, below), we always find a noun or an adjective. After, we must distinguish whether it is a plural noun "كتب" (*kutub*, books) or a singular noun "كتب" (*katb*, writing). In this case, there is no evidence that can confirm that it is one or other of the latter two. Concerning the declination, we can determine the exact declination since we have a grammatical rule that says after a preposition (in this case "تحت" (*tah□ta*, below)), we find a noun in the "مجرور" (*mağruwr* - genitive) case. And with the presence of a personal pronoun "ي" (y) after the noun, tanwin ◌ٍ-(in) will be elimineted. At the end, we find two possibilities: "كُتُبِ" (*kutubi*, books) and "كَتْبِ (*katbi*, writing).

- Heuristics:They are less general rules which are not always true. Among these rules we provoke the following rule.

  Rule: the word "نعم 'preceded by a hyphen" - "means" yes "and not the broken plural of "نعمة" (*niʿma*, grace). Let the following example to which we apply the previous rule.

Example 2:

- نعم لقد خرجت من المنزل

naʿam laqad z□arajtu mina almanzili

Yes I left home.

In this example, the word "نعم" could indicate the broken plural noun "نعم" (*niʿam*, favor) or the adverb "نعم" (*naʿam*, yes) which is used to answer a question with an affirmation. According to the previous rule the word "نعم" is the adverb "نعم" (*naʿam*, yes). But this isn't true in all cases.

By the collect of all groups where broken plural can occur, we can identify the specificity of this plural on each group. Consequently, we can construct the appropriate syntactic or/and morphologic grammar of each group. In the following section, we will see some examples of disambiguation syntactic grammars.

5 Experimentation and evaluation

After the preparation of all syntactic grammar that is necessary for eliminate the ambiguity between broken plurals and the other categories, we experiment our method, by defining several NOOJ syntactic transducers. In what follows we give some examples of these transducers.

Disambiguation grammar of annexation phrase beginning by a number between three and nine

According to the grammar rules, on an annexation phrase (مركب إضافي) the annexed "المضاف إليه" must be in genitive ("مجرور", mağruwr). A phrase beginning with a number between three and nine, is considered as an annexation phrase where the number is the annexing "المضاف" and the rest is the annexed "المضاف إليه". And since annexing "المضاف" is a number between three and nine then a number with the masculine gender shall be followed by a feminine plural noun. While a number with the feminine gender shall be followed by a masculine plural noun.

The following grammar shows the restrictions made on the annexed "المضاف إليه" of an annexation phrase (مركب إضافي) where the annexing "المضاف" is a number between three and nine according to Ellouze (2010).

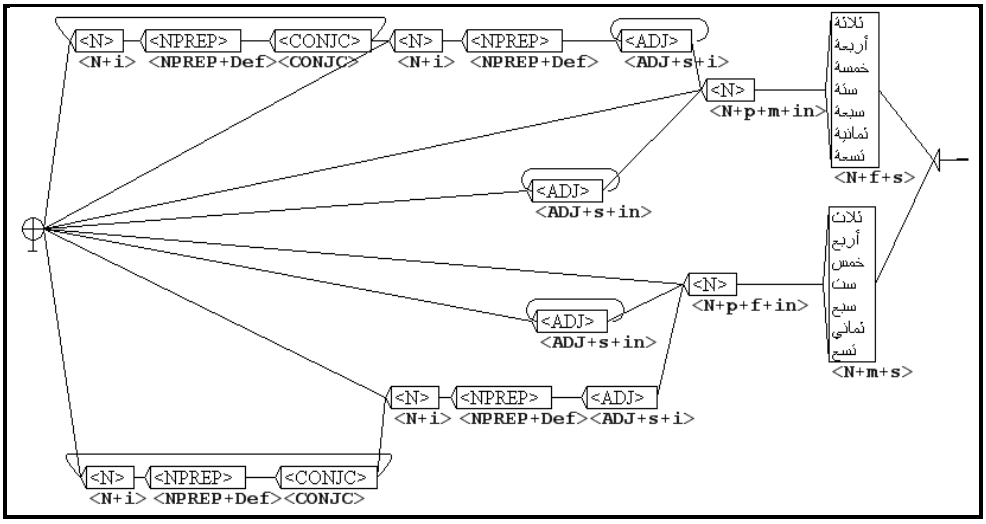


Figure 3. Disambiguation grammar for annexation phrase beginning by number between three and nine

This grammar is able to detect the longest possible solution of disambiguation. When we apply this grammar on a corpus morphologically analyzed, it allows eliminating incorrect interpretations from words of an annexation phrase (مركب إضافي) beginning by number between three and nine. Let seeing the following phrase example.

Example 3:

ثلاثة أقلام جافة صغيرة الحجم

t̲aleṯ̲at ’aqlām ğāfah s□aġirah ālh□ağm

Three dryand small pens.

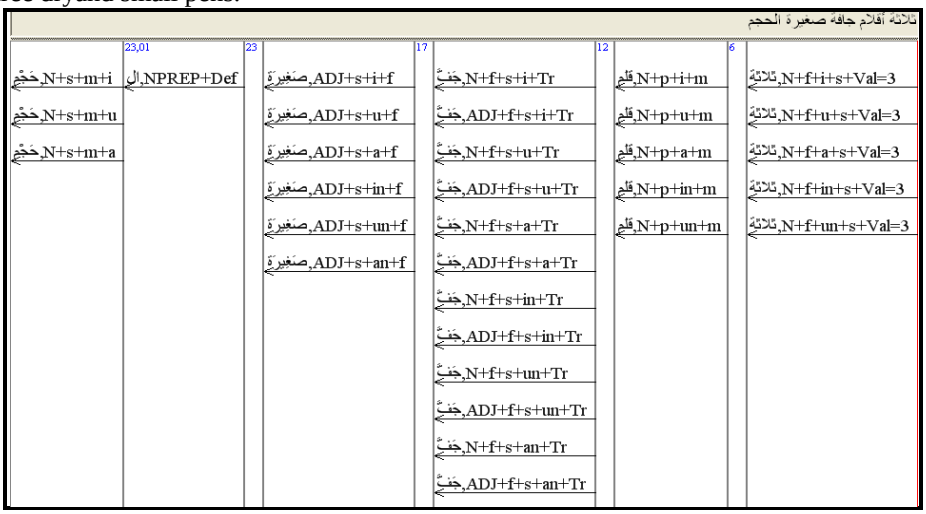


Figure 4. Example of phrase before the application of previous grammar

In figure 4, we show that when we apply morphological analysis on this phrase, we obtain for each word all possible interpretations. For example, the word “جافة” (*ğāfah*, dry) has 12 morphological interpretations. In 6 of these interpretations “جافة” (*ğāfah*, dry) is interpreted as an adjective and in the 6 others as a noun.

But when we apply the grammar of annexation phrase beginning by a number between three and nine, we solve the ambiguity of annexed "المضاف إليه". The declination ambiguity of annexed was resoled thanks to the rules that oblige annexed to have the genitive declination. While the ambiguity of confusion between noun and adjective of the word “جافة” (*ğāfah*, dry) was resolved thanks to the rules that oblige that after noun we find an adjective which describe this noun.

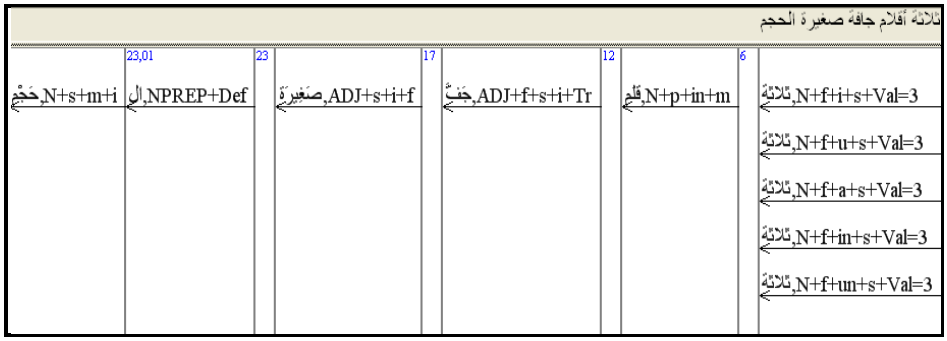


Figure 5. Example of phrase after the application of the previous grammar

In figure 5, we show that ambiguity of annexed "المضاف إليه" was resolved but the ambiguity of annexing "المضاف" was not resolved. To resolve the ambiguity of annexed "المضاف إليه", we chose the following path:

<N+f+s><N+p+s+in><Adj+s+i><Adj+s+i><NPREP+Def><N+i>

The previous path present the longest path that match the phrase in example 3.

After presenting disambiguation grammar of an annexation phrase beginning by number between three and nine, we will present in the following paragraph, disambiguation grammar of an annexation phrase beginning by circumstance.

Disambiguation grammar of an annexation phrase beginning by a circumstance

A phrase beginning by circumstance (e.g., temporary circumstance, place circumstance...) is also considered as an annexation phrase where circumstance is the annexing "المضاف" and the rest is the annexed "المضاف إليه". As previously, on an annexation phrase (مركب إضافي) the annexed "المضاف إليه" must be in genitive ("مجرور" mağruwr). The following grammar shows the restriction made to this phrase.

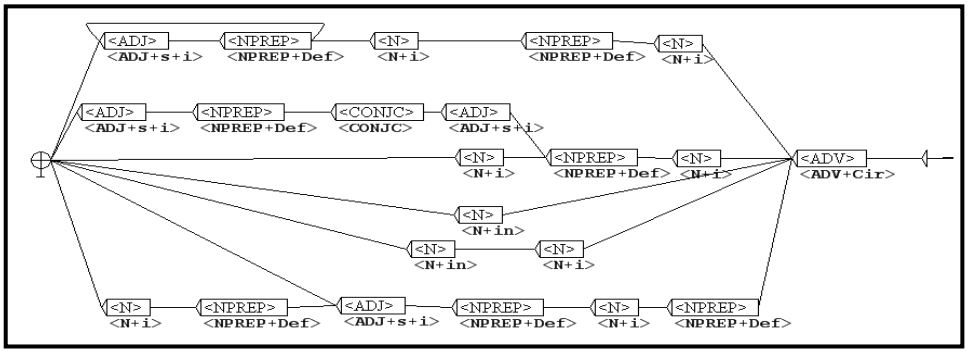


Figure 6. Disambiguation grammar for annexation phrase beginning by circumstance

This grammar is able to detect the longest possible solution of annexation phrase disambiguation. When we apply this grammar on a corpus morphologically analyzed, it

allows eliminating incorrect interpretations from words belonging to the annexed " المضاف إليه". Let seeing the following phrase example.

Example 4:

أمام أبواب المنازل القديمة المغلقة

'amāma 'abwābi ālmanāzili ālqadiymati ālmuġlaqah

In front of doors of old and closed houses

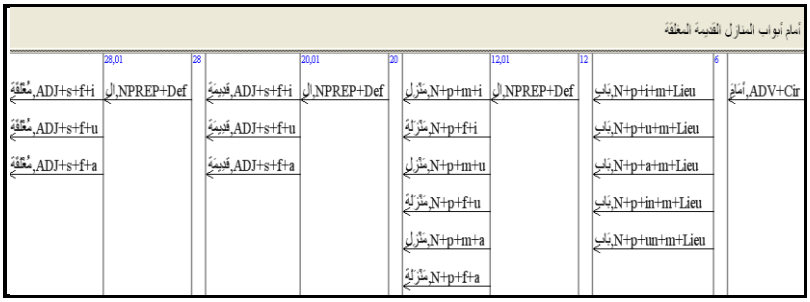


Figure 7. Example of phrase before the application of previous grammar

In figure 7, we show the result of morphological analysis on the phrase of example 4. In this result, we obtain for each word all possible interpretations. For example, the word "منازل" (*manāzil*, houses) has 6 morphological interpretations. In 3 of those interpretations, the word "منازل" (*manāzil*, houses) is interpreted as a plural noun of the masculine singular noun "منزل" (*manzil*, house) and in the 3 others as a plural noun of feminine singular noun "منزلة" (*manzilah*, status).

When we apply previous grammar of annexation phrase beginning by a circumstance, we solve the ambiguity of annexed "المضاف إليه". The declination ambiguity of annexed has been resolved thanks to the grammar rule that oblige annexed to have the genitive declination. While the ambiguity of confusion between the two corresponded singular of the plural "منازل (*Manāzil*, houses) has not been resolved. To rosolve the ambiguity between the two nouns "منزل" (*manzil*, house) and "منزلة" (*manzilah*, status) we must add semantic roles.

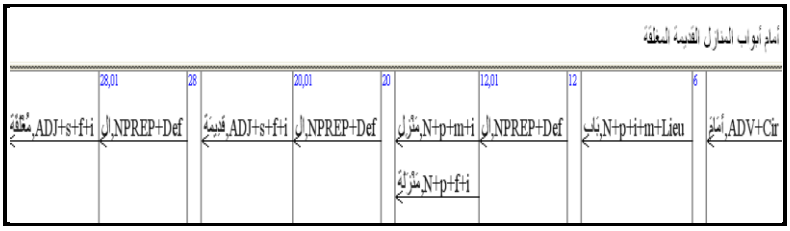


Figure 8. Example of phrase after the application of previous grammar

Figure 8 shows that ambiguity of annexation phrase was partially resolved. Only the correct singular of the plural "منازل (*manāzil*, houses), has not been found. The others words have been correctly matched within any ambiguity. To resolve the ambiguity of annexation phrase, we chose the following path:

<ADV+Cir><N+i><NPREP+Def><Adj+s+i><NPREP+Def><Adj+s+i>

The previous path present the longest path that match the annexation phrase in example 4.

Evaluation

After building the different syntactic grammar of disambiguation, we conducted experimentation on a sample from corpus. This sample contains 2479 words.

Number of BP detected without disambiguation grammars	169
Number of BP detected with disambiguation grammars	160
Number of broken plural correctly detected	122
Number of BP without ambiguity of declination after applying disambiguation grammars	64
Number of BP with ambiguity of declination before applying disambiguation grammars	57

Table 1. The evaluation result of our approach

From the previous table, the ambiguity of BP with other categories is slightly resolved. The number of BP detected was decreased only by 9 words (from 169 to 160) after applying disambiguation grammars. While the number of BP correctly detected is 122. With the application of disambiguation grammars there are 64 of BP are detected without declination ambiguity and 57 of BP are detected with declination ambiguity.

To decrease the ambiguity between BP and other categories, we should include a semantic analysis. And to decrease the ambiguity of declination we should add other morph-syntactic grammars.

6 Conclusion and perspectives

In this paper, we have presented first different methods allowing the disambiguation of broken plurals. Then we have given various ambiguity causes. And finally, we have proposed a linguistic approach for disambiguation. This approach is experimented and evaluated in the linguistic platform NooJ.

There are several perspectives for our work. As a first perspective, we will add more other disambiguation grammars. Also, we should add semantic features to resolve ambiguities of plural words having many singular words with different meaning.

Keywords. Local syntactic grammar, morph-syntactic disambiguation, NooJ, Broken plural.

References

Amardeilh F. 2007. "*Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*", Thesis, Paris-X-Nanterre University.

Al Shamsi F. & Guessoum A. 2006. “A Hidden Markov Model-Based POS Tagger for Arabic”, in *Proceedings of the 8th JADT* (Journées internationales d’Analyse statistique des Données Textuelles), vol 1, pp. 31-41.

Addahdah A., 1992. "معجم قواعد اللغة في جداول و لوحات", librairie Liban, Bairout.

Brill E., 1992. “A simple rule-based part of speech tagger”, In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy,pp. 152-155.

Ellouze S., Haddar K., Abdelwahed A., 2009. “Study and analysis of Arabic broken plural with NooJ”, in *Proceeding of NooJ 2009 conference*, Touzeur, Tunisia.

Ellouze .S, 2010. “*Etude et analyse du pluriel brisé arabe avec la plateforme NooJ*”, Master memory, faculty of economic sciences and management, Sfax, Tunisia.

Freeman A., 2001. *Brill's POS tagger and a Morphology parser for Arabic, PhD*, Department of Near Eastern Studies, Michigan, USA.

Karlsson et al. 1995, "Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text". Mouton de Gruyter.

Khoja S., Garside R. & Knowles G, 2003. “A tagset for the morphosyntactic tagging for Arabic”. In Wilson,A, Rayson, P, McEnery, T (Ed.) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa, Munich, pp.59-72.

Laporte E. & Monceaux A., 1999. "Elimination of lexical ambiguities by grammars. The ELAG system", In *XXII*, Amsterdam-Philadelphie : Benjamins, pp. 341-367.

Mohamed Elhadj Y.O., 2009. “Statistical Part-of-Speech Tagger for Traditional Arabic Texts”, *Journal of Computer Science* No.5 (11), pp. 794-800.

Niama F., 2000. "ملخص قواعد اللغة العربية", editor "*Nahdet Misr for Publishing, Printing and Distribution* ", Cairo, Egypt.

Schmid H., 1994. “Probabilistic part-of-speech tagging using decision trees”, In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, Spain.

Silberztein Max. 2006. *NooJ Manual*. Download from http://www.NooJ4nlp.net

Tlili Guiassa Y., 2006. “Hybrid Method for Tagging Arabic Text”, *Journal of Computer Science*, vol 2 (3), pp. 245-248.

Thorsten B., 2000. “TnT – A Statistical Part-of-Speech Tagger”. In *Proceedings of the Sixth Applied Natural Language Processing Conference* (ANLP). Seattle, WA, pp. 224–231.

Walter D., Jakub Zavrel, Antal van den Bosch & Ko van der Sloot, 2002. “MBT: Memory Based Tagger”, version 1.0, Référence Guide, ILK technical Report-ILK 02-09.

Greek neoclassical compounds and their automatic treatment with Nooj

Gavriilidou Zoe[(1)], Papadopoulou Lena[(2)]

[(1)]*Democritus University of Thrace*
[(2)]*Autonoma Universidad de Barcelona*

1 Introduction

The aim of this paper is to examine a special category of compounds in Modern Greek e.g. *μελισσοκόμος* 'apiculturer', *σαρκοφάγος* 'carnivore', *ανθρωπολόγος* 'anthropologist' (for a study of compounds in *–λόγος*in Greek see Gavriilidou 2006 and in *–logue* in French see Amiot & Dal 2005), often called neoclassical compounds (NC) (Bauer 1983, Lüdelling et al 2001), which are complex words consisting from at least one constituent of savant (i.e. Ancient Greek[1]) origin. Emphasis will be also given to the nature of their constituents, cited in literature (cf. 2.2) as *combining forms*(Marchand 1969, Bauer 1983), *false prefixes*/ *false suffixes* (Lass 1987, Κλαίρης & Μπαμπινώτης 1996) *confixes* (Martinet 1979, Αναστασιάδη-Συμεωνίδη 1986a, 1986b, 1994, Anastassiadis-Symeonidis 1988, Γιαννουλοπούλου 2000), archéoconstituants (Corbin 2001) or *bound stems* (ten Hacken 1994, 2000, Ralli 1988, Ράλλη 2005).

Neoclassical word formation in Greek seems to obey different rules and restrictions than word formation with non savant constituents therefore NCs are idiosyncratic to a certain degree and surely not always transparent from a semantic point of view to non cultivated speakers of Greek. Very often they do not make part of everyday language. On the contrary, they are prolific in scientific terminology. There are some domains, medicine for instance, which make an intense use of them. In this paper, I will try to show how various Nooj applications could facilitate the automatic recognition of such compounds. To do so, I will draw my examples from the vocabulary of medicine.

First I will review the previous literature concerning this kind of composition, then I will describe the phonetic, morphological and semantic characteristics of NCs and of their constituents in Modern Greek and finally, using the vocabulary of medicine as example, I will focus on their automatic treatment with the help of various Nooj applications.

2 Literature review

2.1 Greek NCs

Previous analyses for Greek NCs are made by Αναστασιάδη-Συμεωνίδη (1986a, 1986b, 1994), Anastassiadis-Symeonidis (1988), Ralli (1988), Ράλλη (2005) and Γιαννουλοπούλου (2000).

[1] Neoclassical compounds in other languages may contain constituents of Ancient Greek or Latin origin.

Αναστασιάδη-Συμεωνίδη (1986a, 1994), following Martinet (1979), considers that Greek NCs have been formed in French or English through confixation, which is a mechanism of word formation based on the use of bound stems of Ancient Greek or Latin origin and then entered in Greek vocabulary as loans. She considers that:

- confixes of Modern Greek having their origins to Ancient Greek raise their productivity because of the introduction in Greek of loan words,
- they are used in the creation of new NCs following the rules of learned composition, and
- they preserve sometimes archaic consonant clusters e.g. *δα**κτ**υλογράφος*[2]'dactylographer'.

In order to describe the nature of confixes, she introduced the notion of continuum whose two poles are occupied by (i) stems and (ii) affixes. For her, confixes are between the two poles, closer though to affixes because they are always bound and they belong to close sets.

This position was criticized by Ralli (1988) and Ράλλη(2005) who believes that NCs' constituents are bound stems found closer to free stems than to affixes on the above mentioned continuum because:

i) like stems, they carry a descriptive meaning, while affixes have an instructional or vague meaning,
ii) they carry valency information inherited from the verbal base they have derived from,
iii) they combine with prefixes in word formation e.g $από_{Prefix}$-$πλανώ_{Bound\ Stem}$ 'seduce'. If they were affixes this would mean that we could have words with the structure Affix+Affix. This combination is not allowed in word formation,
iv) they demand a linking vowel when combined with other elements. This is a characteristic of composition,
v) they can be found either in first or in final position e.g. *ανθρωποφάγος*'anthropophage' vs. *μισάνθρωπος* 'misanthrope' and in Greek suffixes can not appear as prefixes as well (cf. also Scalise 1984),
vi) from a synchronic point of view, they can be morphologically and semantically related with other stems from which they derived e.g. –*πλόκος*<*πλέκω*'to knit', -*λόγος*<*λέγω*'to say'.

Γιαννουλοπούλου (2000) like Martinet (1979) considers that NCs are formed via a word formation phenomenon, which lies between compounding and derivation. She adopts the term 'confixation' to describe this phenomenon. For her 'confixes' are morphemes that:

(a) do not coincide with stems, or when they do so, they differ from a semantic point of view e.g. Gr. *ποίηση* 'poetry' vs. Gr.–*ποίηση*'-ization',
(b) come from Classic Greek or Latin elements, and

[2] In Modern Greek there is a preference for dissimilated consonant clusters e.g. χτ [xt] instead of κτ [kt].

(c) belong to the International Scientific Vocabulary, but are also diffused in the general vocabulary (cf. Petrounias 1997).

She claims that the formal approaches (lexicalist and syntacticist position within generative morphology) cannot describe satisfactorily the status of such morphemes because they deny to recognize that the linguistic categories are not separated with clear-cut distinctions. She opts for the theoretical framework of Grammaticalization and argues that confixes are grammaticalized elements because they:

1. carry lexical meaning (like the constituents of the compounds) but they inverse the constituent order in the word, so that the internal word order resembles the constituent order of derivation,
2. are created by secretion,
3. their specific meaning is derived from the "conventionalization of implicature",
4. the diachronic factor plays a significant role in their formation, since they are old morphemes that are re-introduced in the modern language through loan words and neologisms,
5. are found in the International Scientific Vocabulary and in the vocabularies of mass media and politics and therefore their grammaticalization is observed in particularly restricted discourse contexts,
6. can be interpreted as products of a model-word because this function can describe the growing frequency of an item which is ex-free and ex-lexical and
7. are cases of 'incipient grammaticalization' (Hopper 1998) because they are "emergent regularities that have the potential for being instances of grammaticalization" (Hopper 1991).

I will agree with Γιαννουλοπούλου (2000) in that the constituents of NCs are in a phase of grammaticalization which, if not interrupted, will end in their transformation in affixes. I will be based on the notion of *degree of grammaticalization* to support that those constituents are grammaticalized in different degrees. This means that there are constituents, which, for various reasons[3], resist to grammaticalization and are consequently less grammaticalized, especially the ones deriving from verbs and always found in final position. These constituents are closer to the pole occupied by stems in the affix-stem continuum. On the other hand, a higher degree of grammaticalisation is found in constituents deriving from adjectives and nouns e.g. *μικροαμπέρ*'microampere', *μεγαλοκαρδία* 'megalocardia', *καλλιγραφία* 'calligraphy', *γεωλογία*'geology', *κακοφωνία*'cacophony'. Those constituents being mainly in first position in NCs, are closer to the affix pole of the affix-stem continuum. NC' s constituents deriving from Ancient Greek prepositions and adverbs are almost fully grammaticalized e.g. *αμφίβιος*'amphibious'and behave like prefixes (cf. Ράλλη 2005). Therefore, NCs do not form a homogeneous group of compounds in Modern Greek, so that each category requires further investigation.

[3] For example some NC constituents carry valency information inherited from the verb they derived from and this makes grammaticalization more difficult.

2.2 NCs in other languages

NCs in other languages share a number of characteristics with Greek NCs but also differ from them to a certain degree, for instance their constituants come not only form Ancient Greek, as Greek NCs, but from Latin, too. For some linguists (Martinet 1979), NCs in European languages are formed neither with derivation nor with composition but with confixation, which is a third mechanism of word formation with the use of Ancient Greek and Latin constituents. Warren (1990) and Plag (2003) agree in part with the above position arguing that there are differences between compounding and formation of NCs. Plag claims that NCs are distinguished from other types of compounds because they have idiosyncratic formal properties, for instance the presence of a linking vowel, or combinatory and phonological properties. Other scholars (Booij 1992, Scalise 1984, Lüdelling et al 2001) insist on the similarities between NC's formation and compounding. Lüdelling et al (2001) claim that "no clear-cut principled difference can be found between neoclassical and native word formation, because neither phonological properties nor differences in the combinability or in the productivity of these elements allow them to be distinguished from native elements." Amiot & Dal (2005) conclude that not all NC' constituents can be analysed in the same way. They consider that the tools of lexematic morphology are sufficient and suitable for NC's analysis: "these elements can be exponents of LCRs (*micro-* and *-logue*) or suppletive stems of a lexeme, used in constrained contexts (*lud-*, *anthrop-*)". Finally they claim that are easily integrated into the patrimonial lexicon.

The criteria taken into account todescribe NCs constituents in other languages are their bound character, their position in the compound and their semantic and phonological properties.

A number of linguists (Williams 1981, Bauer 1979) consider that the constituents of NCs are affixes, based on the criterion of their boundness. For others (Booij 1992, Plag 2003), they are 'non native roots', but never affixes. For Scalise (1984), they are stems. For Corbin (2001), they are not affixes but archéoconstituants, because affixes have an instructional meaning, while archeoconstituants have a descriptive meaning. For Warren (1990), NCs' constituents are irreducible to the categories of the existing models of word formation, since they are not suffixes (because they have a lexical meaning) nor prefixes.

3 The characteristics of Greek NCs

Greek NCs entered in Modern Greek vocabulary as loanwords directly from Ancient Greek or they were formed in other European languages (mostly in French) and then entered as semi-calques or recognition loan words (emprunts de reconnaissance) (Anastassiadis-Symeonides 1997). In the second case, borrowing of words including neoclassical combining forms from Ancient Greek to French or English permitted the isolation of forms like *anthropo-*, *-logue*, *-morphe*, *cardio-* in those languages and their use in the creation of a large number of scientific terms*ανθρωπολογία*'anthropology', *καρδιοπάθεια*'cardiopathy'. Modern Greek borrowed these terms from French or English or created new terms in analogy with the borrowed ones (Gavriilidou 2006). Consequently, from an etymologic point of view, in Modern Greek vocabulary there are:

i) NCs formed in Ancient Greek and inherited in Modern Greek through their uninterrupted use e.g. *αστρολόγος*[4] 'astrologist',
ii) NCs which are recognition loan words e.g. *τηλέφωνο* 'telephone', that is loan words including two Greek elements, that native speakers recognize as Greek words, thus easily transcribe them with Greek characters,
iii) NCs which are semi-calques, in other words calques formed with at least one Latin constituent, e.g. *κοινωνιολογία*'sociology',
iv) NCs formed in Modern Greek in analogy with the cases ii and iii e.g. *εκλογολόγος*'specialist in analyzing the electoral results'.

Modern Greek NC's constituents display certain characteristics:

i) They only appear as bound constituents of lexemes, even when they receive an inflectional suffix e.g. **-κτόνος*[5] 'who kills'. In that sense they differ from common stems which can exist as autonomous words when assigned an inflection suffix.
ii) According to Ράλλη(2005) they derive from verbs without any derivation affix, whether through conversion (μάχ-$_V$→μαχ-$_N$>-μάχος 'who fights') or through ablaut (πλεκ-$_V$→πλοκ-$_N$> -πλόκος). The initial verbs sometimes no longer exist in Modern Greek,
iii) They can also derive from nouns or adjectives. In that case it is possible to find two allomorphs derived from the same noun or adjective e.g. *μεγάλος>μεγαλ- μεγαλοκεφαλία* 'megalocephalie' vs.*μέγα-μεγάφωνο* 'megaphone', *όνομα>ονοματ- ονοματολογία*'onomatology' vs.*συνωνυμία* 'synonymy',
iv) They sometimes look like free words of Modern Greek, but differ in meaning with them e.g. *–φόρος*[6]'who bears' vs. *φόρος* 'tax',
v) They serve to form terms of scientific or technical field, which are often opaque for native speakers,
vi) They can combine each other e.g. *κεφαλαλγία* 'cephalalgia'; Like free stems, they are linked together with the linking vowel –o,
vii) They bear the characteristic +learned and have the tendency to combine with other constituents bearing the same characteristic (Αναστασιάδη-Συμεωνίδη&Φλιάτουρας 2003) e.g *οστέινος*'osteic' where both constituents are learned vs. **οστέενιος*where the suffix is not learned), *ορειβάτης* 'mountaineer' vs. **βουνοβάτης (*in *βουνοβάτης* the first constituent is not learned), *οδονταλγία* 'odontalgia'vs. **οδοντόπονος* [*odontoache] (where the second constituent is not learned[7]).
viii) The characteristic +learned of the constituents is attributed to the whole compound (Αναστασιάδη-Συμεωνίδη & Φλιάτουρας 2003).

[4] The word *αστρολόγος* was initially used to refer to astrologists and astronomers.
[5] From the Ancient Greek verb *κτείνω* 'to kill'
[6] From the Ancient Greek verb *φέρω* 'to bear' (cf. the combining form of Latin origin *–fer*)
[7] Cf. the non learned synonym *δοντόπονος/ πονόδοντος* 'toothache'.

In the next section, I will exploit the morphological, phonological and semantic features of NCs and of their constituents presented in this paragraph, in order to show the usefulness of linguistic description in NCs' automatic recognition. I will focus in a particular specialised domain, medicine, since, as claimed before (cf. 1), in this science an extremely high number of terms are neoclassical compounds or polylexical terms made of neoclassical compounds.

4 The treatment of Greek NCs with Nooj: the vocabulary of medicine as example

As it has been claimed before, NCs serve to form terms in various special vocabularies. Each special domain selects a particular number of structures and ways of word formation as well as a close number of forms (suffixes, prefixes, stems) to create its terms. In other words, in Greek terminology, a relatively short number of mainly Ancient Greek constituents[8] and sometimes Latin constituents yield to a high number of monolexical specialized terms, which on their turn form polylexical terminological units. For example, medicine vocabulary makes an intense use of:

- neoclassical compounds (as word formation pattern),
- prefixes like *μεγαλο-* 'megalo-', *μικρο-* 'micro-', *υπερ-* 'hyper', *μονο-*'mono', *τετρα-* 'tetra'etc.,
- stems like *αδεν-* 'adeno', *αγγει-* 'angio', *αρτηρι-* 'arterio', *βρογχ-* 'broncho'*καρδι-* 'cardio',*κεφαλ-* 'cephalo'*χειλ-*'cheilo', *χονδρ-*'chondro', etc., usually denoting body parts, or human body organs,
- stems like *παθ-* '-pathy', *αλγ-* '-algia', *δυν-* '-dynia', *λυσ-* '-lysis', *φαγ-* '-phagia', *πληγ-* '--plegia', *πεν-* '-penia'describing a pathogenic situation. These stems are found right before the suffix realization,and in other languages react as suffixes,
- suffixes like *-ίτις/ίτιδα* '-itis', *-εια* '-ia', *-ία* '-ia', *-ωση* 'osis', *-ωμα* '-oma' (for the suffix *-ωμα* '-oma' in Greek medicine terminology cf. Αναστασιάδη-Συμεωνίδη&Φλιάτουρας 2003b)

In fact, medicine specialists use the above mentioned constituents and the model of NCs as formation pattern, in order to make new terms or understand the already existing by segmenting each term in his constituents.

NCs remain unknown words in automatic recognition of texts, especially when specialized texts are parsed, unless they are coded in the dictionary, thing that is costly and time-consuming. There is a need therefore to create local grammars for the automatic recognition of them.

Let us take an example. Terms like *καρδιαλγία* 'cardialgia', *μαστοδυνία* 'mastodynia', *λεμφαδενοπάθεια*'lymphadenopathy' could be recognized, either if they were added in the dictionary, which is a non economic solution, or by reconstructing their formation and representing it with the help of a local grammar.
This is possible to achieve with NooJ, which is a linguistic development environment that includes large-coverage dictionaries and grammars, parses corpora in real time, and

[8] As it is noted above NC constituents belong to close sets of words.

includes tools to create and maintain large-coverage lexical resources, as well as morphological and syntactic grammars. NooJ users can easily develop extractors to identify terms in texts.
In the Greek Nooj module we created a dictionary of NC constituents of medical terms. Each entry is followed by information concerning the meaning of the constituent[9], the class of objects in which it belongs (disease, therapy, medical procedure, etc), its possibility to be a prefix a stem or a suffix, and its equivalent in English:

τετρα,PREF+Maladie+E=tetra
ηπατ,STEM+Maladie+ID1=συκώτι 'liver´+E=hepat
ίτις,SUF+Maladie+ID2=φλεγμονή 'inflamation'+E=itis
ίτιδα,SUF+Maladie+ID2=φλεγμονή 'inflamation'+E=itis

This kind of dictionary in combination with a morphological grammar, where all the words created by a prefix a stem and a suffix are described and inherit the information provided in the dictionary for their constituents, automatically recognize medical terms including NCs' constituents e.g.:

ηπατίτιδα,N+Maladie+ID1= συκώτι +ID2= φλεγμονή

Here follows a morphological grammar for the automatic recognition of terms denoting diseases and including neoclassical elements.

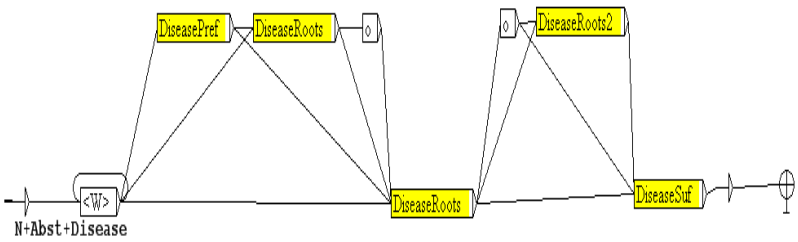


Figure 1. *Morphological grammar of Diseases*

[9] This information is needed for the meaning assignment to the recognized term.

DiseasePref

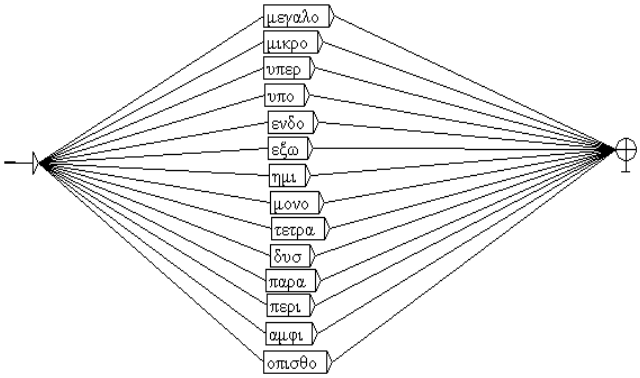


Figure 2. *Diseases' prefixes*

DiseaseRoots2

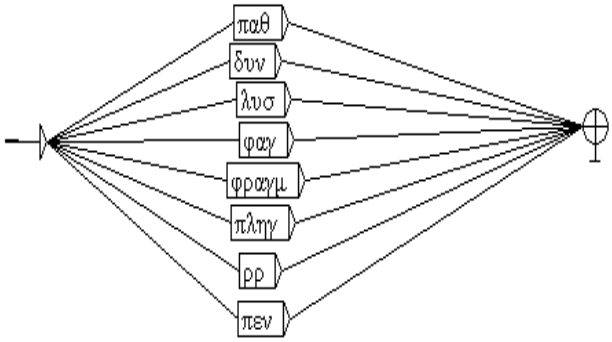


Figure 3. *Stems found before suffix realization*

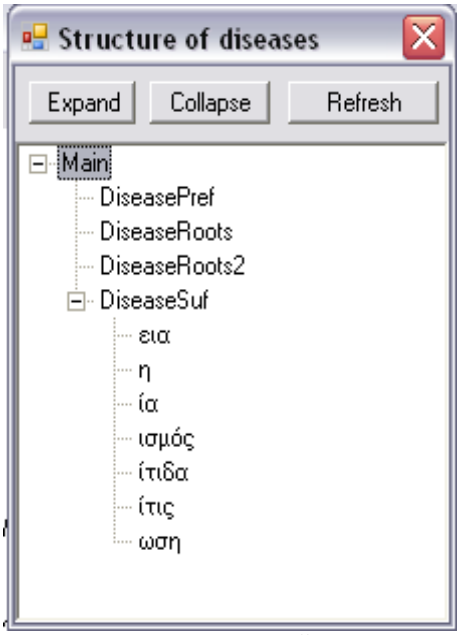


Figure 4. *Disease suffixes*

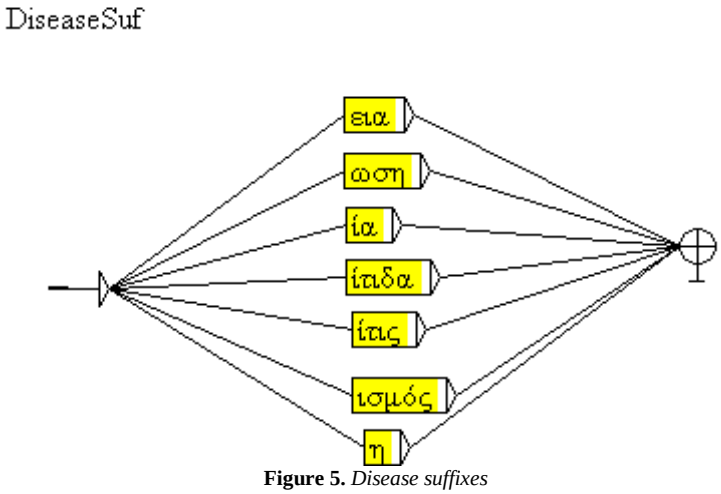


Figure 5. *Disease suffixes*

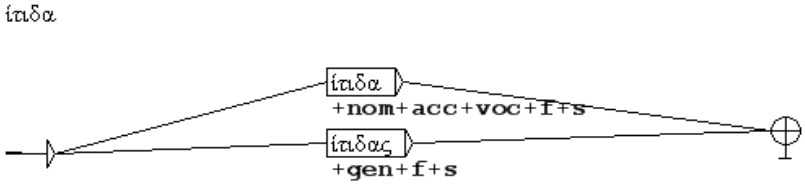


Figure 6. *Graph for the automatic recognition of inflected forms of medical terms in –ίτιδα'-itis'*

As we have seen, we used an example from medicine vocabulary. The same treatment with the help of Nooj could be applied to NCs' constituents found in other special vocabularies or even general vocabulary.

5 Conclusions

In this paper I claimed that NCs do not constitute a homogenous class of compounds in Modern Greek. From an etymologic point of view I separated four different cases of NCs' introduction in Modern Greek vocabulary (from Ancient Greek, from other European languages as recognition loan words or semi-calques, and finally through analogical word formation in Modern Greek). I described NCs and their constituents form a phonetic, morphological and semantic point of view and then implemented this information in Nooj in order to make a morphological grammar and a dictionary of NCs found in medical terms. These two components helped the automatic recognition of medical terms including NCs constituents. The same treatment could automatically recognize NCs' constituents found in other special vocabularies as well.

Key words: neoclassical compounds, bound stems, confixation, grammaticalization, automatic recognition, Nooj,

References

Amiot, D. & B. Dal, 2005. Integrating Neoclassical Combining Forms into aLexeme-Based Morphology, in G. Booij, *et al.* (eds.), *On-line Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5*). Fréjus 15-18 September 2005, University of Bologna, 2007.URL http://mmm.lingue.unibo.it/.

Αναστασιάδη-Συμεωνίδη, Α., 1986a. *Η νεολογία στην Κοινή Νεοελληνική*, Επετηρίς Φιλοσοφικής Σχολής Α.Π.Θ, 65.

Αναστασιάδη-Συμεωνίδη, Α., 1986b. «Η φύση και η παραγωγικότητα του σχηματιστικού στοιχείου –ποιώ», *Μελέτες για την ελληνική γλώσσα*, Θεσσαλονίκη, Κυριακίδης, pp. 261-268.

Anastassiadis-Syméonidis, A., 1988. «La confixation en grec moderne», *Actes du 13ème colloque international de linguistique fonctionnelle*.

Αναστασιάδη-Συμεωνίδη, Α., 1994. *Νεολογικός δανεισμός της νεοελληνικής. Άμεσα δάνεια από τη γαλλική και την αγγλοαμερικανική*, Θεσσαλονίκη.

Αναστασιάδη-Συμεωνίδη, Α., 1996. «Η νεοελληνική σύνθεση», στο *Ζητήματα Νεοελληνικής Γλώσσας. Διδακτική προσέγγιση*, Κατσιμαλή, Γ., Καβουκόπουλος Φ. (επιμ.), Ρέθυμνο, pp. 97-120.

Αναστασιάδη-Συμεωνίδη, Α., 1997. «Διαδικασίες κατά τη δημιουργία των όρων», Πρακτικά 1ου Συνεδρίου ΕΛΕΤΟ «Ελληνική Γλώσσα και ορολογία», Αθήνα, pp. 77-87.

AnastassiadisSyméonidis, A., 2002, *Dictionnaire Inverse du Grec Moderne,* Institut d'Etudes Néohélleniques, Théssalonique.

Αναστασιάδη-Συμεωνίδη, Α, Φλιάτουρας, Α., 2003a. «Η διάκριση [λόγιο] και [λαϊκό] στην ελληνική γλώσσα. Ορισμός και ταξινόμηση», *Πρακτικά 6ου Διεθνούς Συνεδρίου Ελληνικής Γλώσσας*.

Αναστασιάδη-Συμεωνίδη, Α., &Α., Φλιάτουρας, 2003b. Το επίθημα –(ω)μα στην ιατρική ορολογία, *Πρακτικά 4ου Συνεδρίου Ελληνικής Γλώσσας και Ορολογίας*.

Bauer, L., 1983. *English Word Formation*, CambridgeUniversityPress.

Bauer, L., 1998. «Is there a class of neoclassical compounds and if so is it productive?»,*Linguistics*, 36403-421.

Booij, G., 1992. "Compounding in Dutch", *Rivista di Linguistica* 4/1, pp. 37-59.

Corbin, D., 2001. "Préfixes et suffixes : du sens aux catégories." *Journal of French Language Studies* 11/1, pp. 41-69.

Gavriilidou, Z., 2006. Les noms de professions en *–λόγος/-logue* en grec et en français, in (X. Blanco & S. Mejri éds), *Les noms de professions. Approches linguistiques, contrastives et appliquées*, Servei de publicacions, Universidad Autonoma de Barcelona, pp. 128-145.

Γιαννουλοπούλου, Γ., 2000. *Μορφοσημασιολογική σύγκριση παραθημάτων και συμφυμάτων στα νέα ελληνικά και τα ιταλικά*, Θεσσαλονίκη.

ten Hacken, P., 1994. *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*, Hildesheim, Olms.

ten Hacken, P., 2000. "Derivation and Compounding", in Booij, Geert; Lehmann, Christian & Mugdan, Joachim (eds.), *Morphologie - Morphology: Ein Handbuch zur Flexion und Wortbildung - A Handbook on Inflection and Word Formation*, Berlin: Walter de Gruyter, Vol. 1, pp. 349-360.

Hopper P. 1991. "On some principles of grammaticalization". In C.E. Traugott & B. Heine (eds.) *Approaches to Grammaticalization*. Amsterdam: Benjamins, v. I., pp. 17-35.

Hopper P., 1998. "The Paradigm at the End of the Universe", in A. Giacalone-Ramat & P. Hopper (eds), *The limits of Grammaticalization*, Amsterdam: Benjamins, pp. 147-158.

Κλαίρης, Χ., Μπαμπινιώτης, Γ., 1996. *Γραμματική της Νέας Ελληνικής. Δομολειτουργική επικοινωνιακή, Τομ. 1 Το Όνομα της Ν.Ε*, Αθήνα, Ελληνικά Γράμματα.

Lass, R., 1987. *The shape of English*, London, J. M. Dent and Sons.

Lüdeling, A., T., Schmid & S., Kiokpasoglou, 2002. «Neoclassical Word Formation in German». In Booij, Geert; Van Marle, Jaap (ed.) *Yearbook of Morphology, 2001,* pp. 253-283.

Marchand, H., 1969. The categories and types of present-day English word formation. A diachronic-synchronic approach, Beck'sche Verlagsbuchhandlung.

Martinet, A., 1979. *Grammaire fonctionnelle du français*, Didier, Paris.

Petrounias, Ev., 1997. "Loan Translations and the Etymologies of Modern Greek", in *Greek Linguistics 95*. Salzburg: University of Salzburg, pp. 791-801.

Plag, I., 2003. *Word-Formation in English*. Cambridge: Cambridge University Press.

Ralli, A., 1988. *Elements de la morphologie du grec moderne. La structure du verbe*, Phd Diss. Université de Montréal.

Ράλλη, Α., 2005. *Μορφολογία*, Πατάκης, Αθήνα.

Scalise, S., 1984. *Generative Morphology*, Dordrecht (Holland) / Cinnaminson (U.S.A.), Foris Publications.

Warren, B., 1990. "The Importance of Combining Forms." In W. Dressler *et al.* (eds), *Contemporary Morphology*. Berlin / New York, Mouton de Gruyter, pp. 111-132.

Williams, E., 1981. "On the Notions 'Lexically Related' and 'Head of a Word'', *Linguistic Inquiry* 12, pp. 245-274.

Deriving Adjectives and Nouns from Numerals

Kristina Vučković[(1)], Sara Librenjak[(1)], Zdravko Dovedan Han[(1)]

[(1)]*Department of information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb*
Zagreb, Croatia.

Abstract

The paper discusses formation of nouns and adjectives from numerals in Croatian language using NooJ morphological grammars. Description of this semantic group using productive morphology in NooJ is important in order to minimize the number of unnecessary entries to the dictionary, thus saving on space and time.

We believe that detailed description of numeral nouns but also numerals behaving as prefixes to adjectives will increase both precision and recall in recognition of noun phrases <NP> and prepositional phrases <PP>. Our results will be demonstrated by reapplying the local grammars for recognition of <NP> and <PP> chunks to the corpus.

1 Introduction

That grammarians and lexicographers can not quite agree upon different aspects of numeric types of words is probably very common thing regardless the language (Hurford, 2003). Croatian is not much different (Franić, 2008).

Numeral nouns in Croatian language, as well as in many other languages, are considered a specific group of words. This is mainly due to the fact that this group possesses a limited number of semantic roots (numeral part) and uses the finite number of possible suffixes (Barić et al., 2005). It is similar with the numeral adjectives.

In Croatian texts we find four different types of numerals: numerals as numbers like 1, 7, 10, 150, 145 678 (Bekavac, 2005; Vučković, 2009; Vučković et al., 2010), numerals as words like *jedan* (one), *tri* (three), *pet* (five) (Bekavac, 2005; Vučković, 2009; Vučković et al., 2010), numerals as prefixes to adjectives (***jedno****dnevni* (one day old), ***tro****mjesečni* (three months old), ***peto****godišnji* (five years old), ***šest****kilometarski* (six kilometers long), ***dvo****litarski* (two liters weight)) and numerals as nouns (***dvadeset****ica*, ***dvadeset****orka*, ***dvadeset****ak*, ***dvadeset****ina)*.

In this work we will concentrate on the last two types of numerals' usage and explain different ending possibilities in more details in sections two and three respectively. We will conclude this paper with the evaluation of our model presenting some individual measures for these two types of derivations and also comparative measures for noun phrases <NP> and prepositional phrases <PP> recognition since both phrases have adjectives and nouns as building blocks.

2 Deriving Adjectives

One morphological (see Figure 2) and one syntactic grammar (see Figure 1) had to be built for the derivation of adjectives with numerals as prefixes. The reason for the syntactic

grammar lies in the fact that there are occurrences in the text where numeral preceding the adjective can be separated from the adjective with a dash "-". Since NooJ recognizes a dash as a delimiter and word form as a sequence of letters between two delimiters (Silberztein, 2003), so far, it does not recognize strings like Digit_Delimeter_WordForm as one token but rather as three on the lexical level.

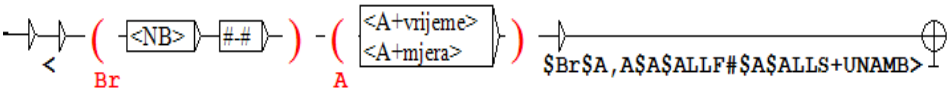


Figure 1. Syntactic grammar for recognition of numeric adjectives

The grammar in Figure 1 recognizes any number <NB> immediately followed by a dash <#-#> immediately followed by an adjective of time <A+vrijeme> or measure <A+mjera>. Recognized string is annotated as an adjective with all the syntactic and morphological characteristics of an adjective that it is made of. For example:

3.2 28-godišnji (28 years old)

10-minutni (10 minutes long)

90-dnevne (90 days old)

The grammar in Figure 2 recognizes adjective made of a numeral (written with letters not digits) and any string of letters recognized as an adjective that is in the dictionary marked as an aadjective of time <$A=:A+vrijeme> or measure <$A=:A+mjera>. Again, the recognized string carries the morphological and syntactic annotations of its adjective.

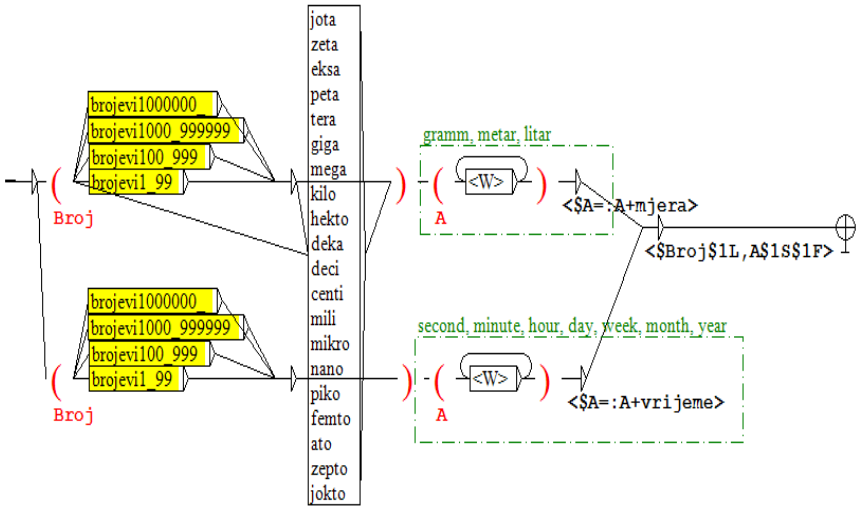


Figure 2. Morphological grammar for recognition of numeric adjectives

On the path that recognizes adjectives of measure, there is an additional node with prefixes that denote greater or smaller measures than the base ones (meter, liter, gram). This node can be preceded with an additional number as well.

4.2 dekagramski uteg	(decagramic weight)
hektolitarski stupanj	(hectoliteric degree)
kilometarski red	(kilometric cue)
devedesetmilimetarski projektil	(ninetymilimeteric missile)

After adding these grammars we were able to remove 148 adjectives(130 were concerning the time and 18 concerning other measures) from the dictionary that were covering only a small portion of adjectives with numerals as prefixes anyhow. In this way, we were able to reduce the size of the dictionary and at the same time raise the number of recognized forms.

Similar grammars can be built for other adjectives that take a numeral as prefix, such as:

5.2 ***dvo****dimenzionalan,* ***dvo****krevetan,* ***dvo****djelan,* ***dvo****član,* ***dvo****krak,* ***dvo****kutan,* ***dvo****znamenkast, …*

- two dimensional, double bed, of two parts, of two members, of two sidesed, of two angles, of two digits, …

tro*dimenzionalan,* ***tro****krevetan,* ***tro****djelan,* ***tro****član,* ***tro****krak,* ***tro****kutan,* ***tro****znamenkast, …*

- three dimensional, triple bed, of three parts, of three members, of three sides, of three angles, of three digits, …

But, we leave this for some future work.

3 Deriving Nouns

Numerals as nouns are built from a numeral part and any of the following suffixes: -ica, -jica, -orka, -ojka, -ak, -ina. Some numerals require to undergo certain changes before the productive suffix is applied, like numerals ‘*dva*’ (two), ‘*tri*’ (three), ‘*četiri*’ (four), while others remain in the same form like ‘*pet*’ (five), ‘*šest*’ (six), ‘*devet*’ (nine).

However, these changes do not affect the same numbers in the same way if different suffix is applied. We will now show these changes in more details.

3.1 Counting Masculine Nouns with -ICA

When talking about ‘two boys’, it is perfectly proper to say ‘*dva dječaka*’ and use the numeral ‘*dva*’ to denote the number of boys. However, we can use the numeral noun ‘*dvojica*’ for the same purpose:

6.2 2 boys -> dvojica dječaka	7 boys ->*sedmorica dječaka*
3 boys ->*trojica dječaka*	8 boys ->*osmorica dječaka*
4 boys ->*četvorica dječaka*	9 boys ->*devetorica dječaka*
5 boys ->*petorica dječaka*	10 boys ->*desetorica dječaka*
6 boys ->*šestorica dječaka*	

The morphological grammar shown in Figure 3 gives a detailed view how to build numerical nouns with the suffix **-ica**. The grammar uses four subgraphs: **jedinice** for recognizing numerals from 2 to 9 (see Figure 4); **desetice** for recognizing numerals like 20, 30, 40, etc. (see Figure 5); **11to19** for recognizing numerals 11 through 19 (see Figure 6); **stotice** for recognizing numerals like 100, 200, 300, etc (see Figure 7). Thus, this grammar builds all forms of numeric nouns from numerals 1 to 999 in seven different cases (Nominative, Genitive, Dative, Accusative, Vocative, Locative, and Instrumental) with possible double endings for Dative, Instrumental and Locative. All the recognized words are also marked as nouns <N> of numeric type <+Type=bi> in masculine gender <+Gender=m> and in plural <+Nb=p>.

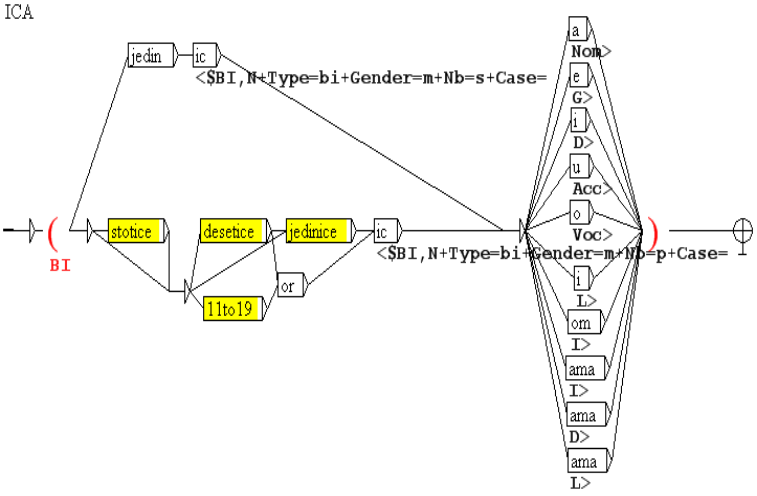


Figure 3. Building numeral nouns with - ica suffix (1st usage)

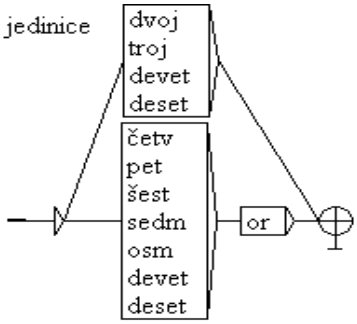


Figure 4. Subgraph for numerals 1 to 10

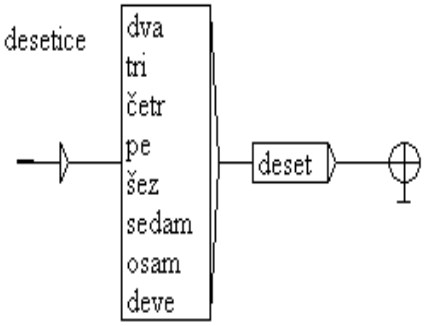


Figure 5. Subgraph for numerals 20, 30, etc.

On the syntactic level, it is important to notice the following characteristics. When using the numeral '*dva*' we have one <NP> which consists of one main noun '*dječaka*' that is proceeded by a numeral denoting the main noun '*dva*', i.e. <$_{NP}$*dva dječaka*>.

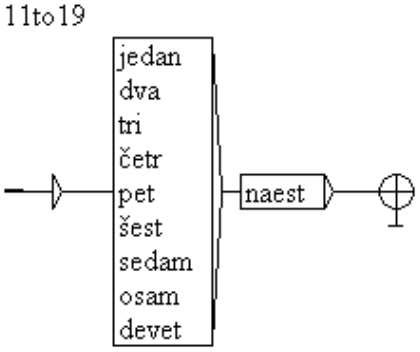


Figure 6. Subgraph for numerals 11 to 19

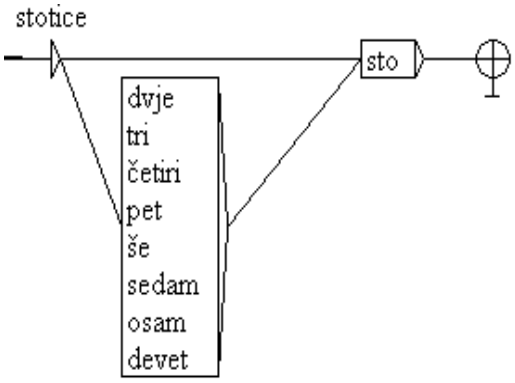


Figure 7. Subgraph for numerals 100, 200, 300, etc.

However, if using numeral noun '*dvojica*' for the same purpose, than we have two <NP>s. The first is <$_{NP}$*dvojica*>and the second is <$_{NP}$*dječaka*> where the second <NP> is always a noun in Genitive case, masculine and plural form and is an Attribute to the first <NP>. So the full notation for our example is <$_{NP+Nom+f+p}$ dvojica <$_{NP+Attribute+G+m+p}$ dječaka>>

Although it is quite usual that numeral noun of this type will be followed by another noun in genitive case, this is not an obligatory scenario and they may occur all by themselves.

Dvojica *su stigla jučer.* (**Two of them** came yesterday.)

Dvojica *dječaka su stigla jučer.* (**Two** boys came yesterday.)

3.2 Other meanings of suffix -ICA

The usage of suffix **-ica** as explained in the previous chapter is not its only usage. This ending is also used when talking about the size of clothes, grade in school (numerals 1 through 5) or money bills (10, 50, 100, 500). In addition to the plural forms, numeral nouns derived using the grammar in Figure 8 build the singular forms as well and they are all in feminine gender.

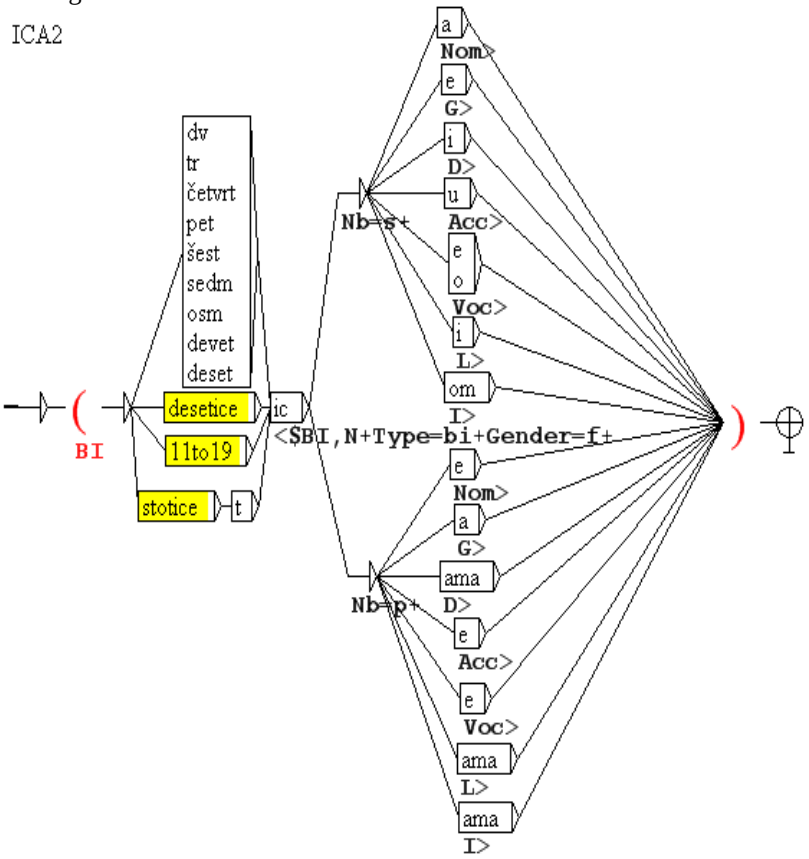


Figure 8. Second grammar for suffix -ICA

The grammar in Figure 8 uses the same subgraphs for numerals 11 through 19 (Figure 6), 20..90 (Figure 5), 100..900 (Figure 7) as the grammar in Figure 3.

3.3 Approximate Number with - AK

Numbers from 9 through 20, and numbers ending in one or two zeros, like 20, 30, 400, 500, may take a suffix **-ak** when talking about approximation:

*devet**ak** dječaka* (**aroundnine** boys)

*stotinj**ak** djevojčica* (**aroundhundred** girls)

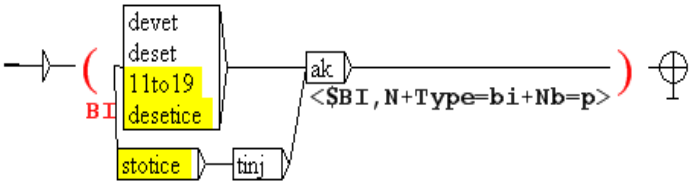


Figure 9. Numeral nouns with suffix -ak

For the syntactic level of analysis it is important to notice that this form of numeral noun is usually followed by another noun in genitive, regardless the gender.

It is also possible to find a preposition in front of it. In this case, the numeral noun ending in **-ak** is in Accusative case.

s<$_{\text{N+bi+Acc+p}}$ desetak> prijatelja (**with**<around ten> friends)

nakon<$_{\text{N+bi+Acc+p}}$ pedesetak> godina (**after**<around fifty> years)

3.4 Fractions with suffix -IN(K)A and -NINA

When talking about fractions in Croatian, we can talk about two types of fractions: fractions of time and all other fractions. The difference between these two is the suffix the numeral takes in order to build a numeric noun. Thus, when taking of fractions of time the suffix **-inka** is used.

*treć**inka** vremena* **(**third of time)

*pet**inka** vremena* (fifth of time)

If the suffix **-ina** is used, than we are definitely not talking about the time but some other type of fraction, or part of something.

*treć**ina** dječaka* (third of boys)

*pet**ina** djevojčica* (fifth of girls)

Numeral nouns determining fraction that are derived from numerals 11 through 19, billion, another word for thousand (hr. *hiljada*) as well as the word for hundred (hr. s*to*) use the suffix **-nin**.

Regardless the suffix, all the numeral nouns defining the fraction are in feminine gender (see Figure 10). The grammar uses subgraphs **desetice** (see Figure 5) and **11 to 19** (see Figure 6) in the same manner as the previously explained grammars.

For the syntactic analysis it is important to notice the following occurrences.

If there is no other numeral or if the numeral preceding the fraction part is *one*, than the numeral noun is in Nominative case and singular:

<N+bi+f+s***trećina***>*dječaka* (<**third**> of boys)

jedna<N+bi+f+Nom+s***trećina***>*dječaka* (one <**third**> of boys).

However, if the number that precedes the numeric noun is two, three or four, than the numeric noun is in Nominative case and plural:

dvije< N+bi+f+Nom+p***trećine***>*dječaka* (two <**thirds**> of boys).

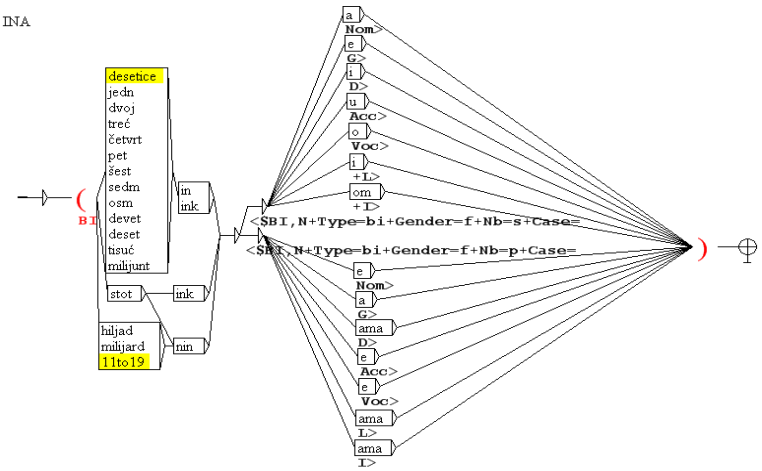


Figure 10. Grammar for building numeric nouns for fractions

Finally, if the number that precedes the numeric noun is five or higher, than the numeric noun is in Genitive case and plural:

pet< N+bi+f+G+p***trećina***>*dječaka* (five <**thirds**> of boys).

Case is marked according to the final suffix following the suffix **-in(k)** or **-nin** (see Figure 10).

3.5 Different meanings of -OJKA and -ORKA

The final type of the suffix used for building numeral nouns in Croatian is the ending **-ojka** or **-orka** as shown in Figure 11.

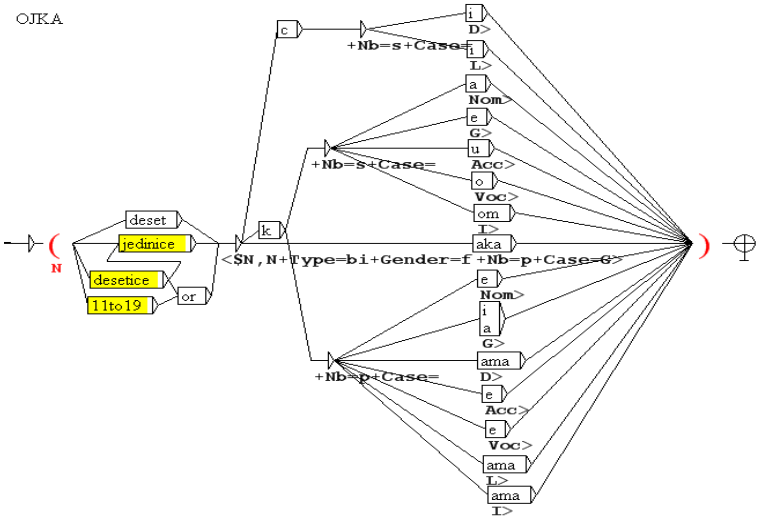


Figure 11. Main graph for suffix OJKA / ORKA

This grammar uses three subgraphs **jedinice** (see Figure 4), **desetice** (see Figure 5) and **11to19** (see Figure 6). All of the nouns <N> derived in such a manner are of Type numeral <+Type=bi>, feminine in gender <+Gender=f> and have both singular and plural endings for all seven cases.

This group of numeral nouns denotes group of any number of people, for example:

Petorka *je stigla.* (**A group of 5 people** has arrived.)

Tridesetorka *je pjevala.* (A **group of 30 people** was singing.)

However, nouns derived from numerals 2, 3 and 4 are also used when talking about a grade in school:

Dobio sam ***dvojku*** *na testu.* (I got a **D (two)** on the test.)

Nouns formed with suffixes **-ojka** or **-orka** are also used when talking about multiple births or bus/tram numbers:

Dobili smo ***dvojke***. (We had **twins**.)

Rodila je ***četvorke***. (She gave birth to **quadruplets**.)

Na posao idem ***trojkom***. (I take tram **number 3** to work.)

4 Results

The samples were collected in two manners, by using the data from Croatian national corpus (HNL), and web search for the lexical forms not present in the corpus. The sample contained 454 sentences which were selected on the basis of the following criteria:

- each sentence had to contain at least one form of the numeral relevant to the research,

- all grammatical forms (case and number) had to be represented in the sample for numerals between 1 and 10, and few representative ones for the larger numbers which follow the same grammatical pattern,
- the numeral form had to correspond semantically to the research; sentences containing homographs with content unrelated to the numeral's meaning (as explained in previous chapters) were eliminated,
- sentences had to be real utterances of Croatian language, and were chosen regardless of topic and register.

After all the relevant forms were exhaustively represented by the sample, sentences were marked morphologically and grammars were tested using NooJ. Seeing as they were all recognized and marked correctly, the grammars were shown to be successful in processing Croatian adjectives and numerals built from nouns.

4.1 Individual results for adjectives

The morphological and syntactic grammars for adjectives built from numerals as prefixes were tested on the previously described sample sentences in which 166 sentences consisted of at least one adjective marked with <+time> or <+measure> annotation. Grammars have recognized all 149 occurrences of adjectives, and annotated correctly 30 adjectives of measure and 119 adjectives of time. An overall F1 measure for this model is 1 (P=1, R=1).

4.2 Individual results for nouns

The morphological grammar for numeral nouns was tested on 288 examples all of which were recognized and marked correctly thus giving us also all three measures (precision, recall and F1-measure) of 1.

4.3 Comparative results for <NP> and <PP>

Both previously described text samples were tested for <NP> and <PP> detection twofold i.e. without and with the new grammars included in the testing. The results of their performance in terms of their precision P, recall R and F1-measere are given in Table 1.

New grammars		**Not applied**	**Applied**
<NP>	**P**	0,765	**0,839**
	R	0,898	**1**
	F1	0,826	**0,912**
<PP>	**P**	0,925	**0,995**
	R	0,804	**1**
	F1	0,86	**0,997**

Table 1. Standard measures for <NP> and <PP> recognition of sample sentences

The model using the new grammars outperforms the model without the new grammars with recall of 1 for both <NP> and <PP> detection.

5 Conclusion

The article describes the morphological grammar for building nouns and morphological and syntactic grammars for building adjectives from numerals. Although all of these adjectives and nouns could be noted in the NooJ dictionary as individual entries, this would be not only expensive in time and space, but above all unnecessary due to the morphological and syntactic RTNs within NooJ (Silberztein, 2003) which allows us to recognize the word in the text but also to connect it with the necessary annotation that would otherwise be noted via dictionary and inflectional/derivational grammars.

As the results show, all of the grammars perform quite well raising the F1 measure for detection of <NP> and <PP> chunks from 0,912 to 0,997. From this we can conclude that the entire model for chunking and parsing Croatian texts is further improved which is, after all, our final goal.

Key Words
Derivation, numeral nouns, numeral adjectives, Croatian, morphological grammars, syntactic grammars, NooJ.

Acknowledgments

This work was done within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grant 130-1300646-1776.

References

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević V. and Znika, M. 2005. *Hrvatska gramatika*, Školska knjiga, Zagreb.

Bekavac, B. 2005. *Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima*. (Computer recognition of named entities in the contemporary Croatian texts.) PhD Thesis, Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Hurford, J. 2003. "The interaction between numerals and nouns". In F. Plank (ed) *Noun Phrase Structure in the Languages of Europe*, Volume 20-7 of Empirical Approaches to Language Typology, 561-620.

Franić, I. 2008. "Leksikografski status brojevnih riječi u Rječniku hrvatskoga kajkavskoga književnog jezika" (Lexicographic Status of Numbers in the Dictionary of Croatian Kajkavian Litterary Language) In *Rasprave Institute za hrvatski jezik i jezikoslovlje*,107-131.

Silberztein, M. 2003. *NooJ Manual*, available at the web site http://nooj4nlp.net (200 pages).

Vučković, K. 2009. *Model parsera za hrvatski jezik*. (Model of a Parser for Croatian Language) PhD Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, 2009.

Vučković, K., Tadić, M., Bekavac, B. 2010. "Croatian Language Resources for NooJ". In V. Lužar-Stiffler, I. Jarec, Z. Bekić (eds) *Proceedings of the 32nd International Conference on Information Technology Interfaces*, SRCE University Computer Centre, University of Zagreb, Zagreb, 121-126.

Version 4 Greek NooJ Module: Adverbs, acronyms and words with Latin characters

Papadopoulou Lena[(1)], Chatzipapa Elina[(2)]

[(1)] *Autonomous University of Barcelona, lepapad@hotmail.com*
[(2)] *Democritus University of Thrace, elinaxp@hotmail.com*

1 Introduction

The main objective of this paper is to present the ongoing work on the Modern Greek NooJ Module (Gavriilidou, Papadopoulou, & Chatzipapa, 2008); (Gavriilidou, Papadopoulou, & Chatzipapa, to appear); (Papadopoulou & Gavriilidou, 2010), which aims to achieve the best possible automatic processing of Greek language. Within the scope of this aim, tests of our data had to be constantly performed, which allowed possible lacks and deficiencies to be brought to the surface and to be further fulfilled.

Linguistic analyses of our corpora have demonstrated that the great majority of the unknown word forms are adverbs, abbreviations and words written with Latin characters. In order to reduce the number of unknown words, and improve, in this way, text annotation of the Greek NooJ Module we put forward the construction of : a) an adverb dictionary and supplementary syntactic grammars, b) a dictionary of acronyms and c) a dictionary of lexical units written with Latin characters that occur with high frequency in Greek texts.

2 Adverbs

Our lexicographic work on adverbs is placed within the framework of M. Gross (1986), according to whom both simple (*σύντομα/briefly*) and multiword adverbs (*μια για πάντα/once and for all*) have to be treated as a single lexical unit, which dispose its own syntactic and semantic patterns (Blanco & Guenthner, Multi-lexemic Expressions: An Overview, 2004). Given that the present work constitutes a first step towards a systematic treatment of the adverbs with the platform of NooJ, our lexicographic work will follow the theory of *Lexicon-Grammar* of M. Gross regarding the nomenclature in combination with the semantic criteria that traditional grammar define for adverbs.

By the term adverb we refer to these structures that have adverbial value and they could be simple adverbs (*αύριο/tomorrow*), derived adverbs (*γρήγορα/quickly*), circumstantial complements (μεταπόδια/on foot) or circumstantial subordinate clauses (*μέχριναπεις κύμινο /like a house on fire*). All the aforementioned types of adverbs are included in our adverb dictionary. More particularly, our macrostructure consists of 4.000 simple and 1.400 multiword adverbs.

Main source for the collection of the simple adverbs was the network www.komvos.edu.gr and, especially, the electronic version of *Modern Greek Language Dictionary* of M. Triantafyllidis and the *Reverse Index* of A. Anastasiadi-Simeonidi. The latter constituted an extremely useful tool, as it allowed as to perform quests according to typical adverb suffixes, such as *–ως/-ώς*, *-ιστί* and *-ηδόν*. Although the lexicographic elaboration of simple adverbs had been normally carried out, homography problems arose

during the linguistic analysis of our texts, concerningadverbs ending to *-α* (*σύντομα/briefly* and *brief*) and *-ά* (*αργά/slow* and *slowly*), and the singular feminine nominative/accusative or vocative form or plural neutral nominativeaccusative or vocative form of several adjectives (e.g παραμυθένια).

This problem was solved by the construction of a general syntactic grammar (Figure 1), according to which when an adverbial (ADV) form follows a verb (V) and is followed by an article (AR) or a preposition (PREP) it will be annotated correctly as an adverb (Figure 2) and not as an adjective (A), as it would be annotated before (Figure 3), without the aid of this grammar.

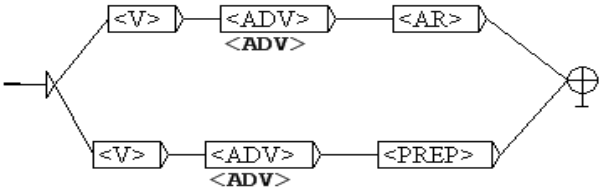


Figure 1. Syntactic grammar for the disambiguation of adverbs and adjective forms

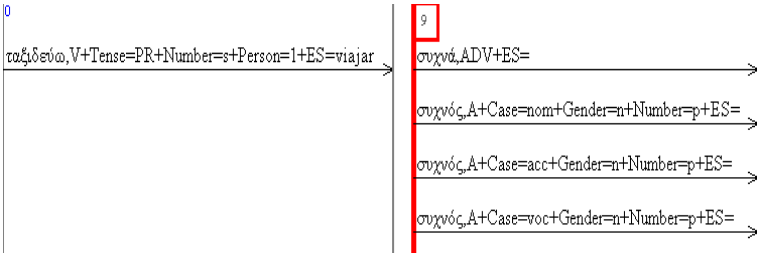


Figure 2. Annotation without the syntactic grammar

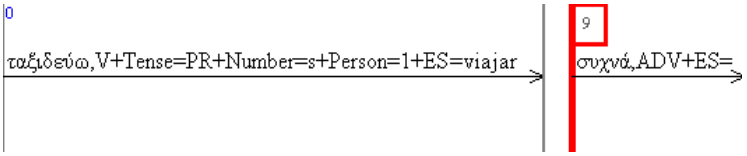


Figure 3. Annotation with the syntactic grammar

Regarding the compound adverbs, we get free and frozen adverbial expressions. The first type concerns free phrasemes (Mel'cuk, 1998), whose meaning is the sum of the meanings of their component elements ("A" ⊕ "B" = "A ⊕ B"). Representative examples of such adverbial expressions are the structures of the dates (*Στις 27 Μαρτίου 2010/On Saturday 27 March 2010*) and time expression (*Στις πέντε η ώρα/At five o'clock*). Although compositional adverbials are characterized by freedom, we studied and formalized a series of such sequences in a syntactic grammar (Figure 4) in order to be

recognized as a whole, ameliorating in this way the text annotation. For example, the phrase «*τηνπροηγούμενηΚυριακή»* (*last Sunday*), which would be simply analyzed as Article+Adjective+$Noun_{time}$, is now annotated as a temporal adverb.

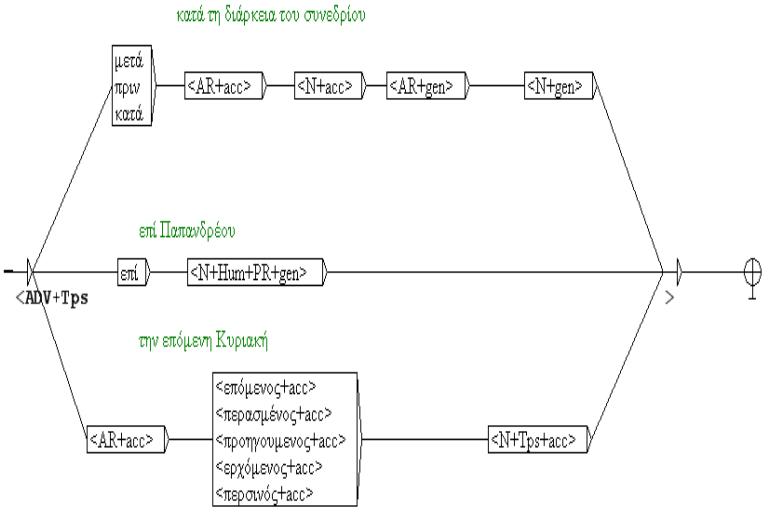


Figure 4. Sample of the syntactic grammar of free adverbials

The second type of adverbial expressions belongs to the quasi-phrasemes and full phrasemes. This means that their meaning is not compositional, as they present opacity in semantic, syntactic and morphological level ("A" ⊕□"B" = "A ⊕□B⊕□C" or "A" ⊕□"B" = "C"). However, they do have a syntactic behavior that is homologous to simple adverbs. For example, although the compound adverb *παρά τρίχα*[1] does not call to mind its meaning (*by a whisker*), it is considered syntactically an adverb. We have to mention that the degree of semantic transparency varies, as Baptista (2004) notes. Gross (1996) also refers to the semantic frozenness and to the syntactic frozenness. For example, the compound adverb *μεταίδιαμουταμάτια (with my own eyes)* presents low semantic opacity, as we can easy understand its meaning, and its syntactic pattern contains in a restrictive way variable elements (the possessive pronoun *μου/ my*). Studying the structure of such compound adverbs we formalized their structure in a syntactic grammar (Figure 5).In this way, all the discontinuous forms of our frozen adverbial expressions (i.e. *μεταίδιασουταμάτια/with your own eyes*, *μεταίδιατουςταμάτια/with their own eyes*) are recognized and annotated as a unit by NooJ (Silberztein, 2008).

[1] The literally meaning of *παρά τρίχα* is **from hair*.

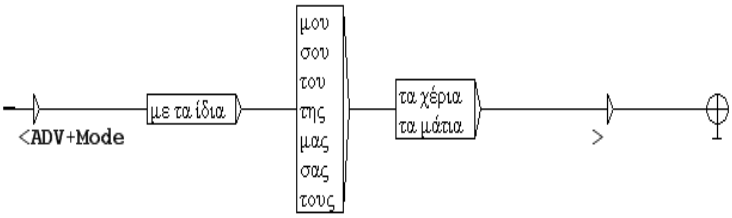


Figure 5. Sample of the syntactic grammar for compound adverbs

With respect to the microstructure of this dictionary, we followed the format of the Greek NooJ dictionary. On the one hand, we expanded the macrostructure of the existent dictionary with more adverbs, as we worked on the homography between the adverbs and adjectives´forms (cf. above). On the other hand, we amplified the microstructure by the introduction of the eight semantic categories of the adverbs, which are defined typically by answering questions with interrogative adverbs (i.e. *πώς;/how, πότε;/when?, πού;/where?*, etc), : i) mode, ii) place, iii) time, iv) quantity, v) negation, vi) doubt, vii) affirmation (Triantafyllidis, 1977).

3 Acronyms

The Greek NooJ dictionary was also grew up by the introduction of acronyms into its macrostructure. With acronyms we refer to abbreviations of the written form of the language that have been transformed into a unit of the spoken form (Anastasiadis-Simeonidis, 1986) or more simply initialisms that have become independent lexical units (Xydopoulos & Vazou, 2007), such as *OHE/UNO* (*Οργανισμός Ηνωμένων Εθνών/ United Nations Organization*).

Acronyms have been treated as lexical units. On the one hand, they pose their own semantics, as they can be polysemous, i.e. the *ΔΕΗ* (*Δημόσια Επιχείρηση Ηλεκτρισμού/Public Power Corporation*) correspond both to the respective company and to the bill of that company, or they can present homonymy, i.e. EOK correspond to *Ευρωπαϊκή Οικονομική Κοινότητα /European Economic Community*, *Ελληνική Ομοσπονδία Καλαθοσφαίρισης/Hellenic Basketball Clubs Association* and to *Εθνικός Οργανισμός Καπνού/National Organization of Tobacco*. On the other hand, they have their own morphosyntactic status, having usually nominal function in syntax and carrying grammatical characteristics, which sometimes are independently formed of the acronym head, i.e. MAT (*Μονάδες Αποκατάστασης Τάξης /Units for the Reinstatement of Order*) behave as neutral in a sentence (*πλακώσανε τα ΜΑΤ/Units for the Reinstatement of Order* arrived unexpectedly), whereas it is feminine.

The macrostructure of our dictionary contains in total 400 acronyms and the microstructure is organized in four fields. First, we have provided the orthographic variants of our lemmas, i.e. both *ΟΤΕ* and *O.T.E.* are provided (*Οργανισμός Τηλεπικοινωνιών Ελλάδος/ Hellenic Organization of Telecommunications*). In the second field, we have assigned the part of speech to which belongs each lemma. We have to mention, in this point, that all the abbreviations that we have treated pose nominal function. In the following field, we have annotated their morphological properties, that is the gender and the number in which we meet them. For example, the ATE (*Αγροτική Τράπεζα Ελλάδος/National Bank*

of Agriculture) corresponds to the morphological code *INDECfs*, which means that we meet it only in feminine of the singular. However, during the assignment of the grammatical characteristics, problems of the gender definition raised. As we have aforementioned, there are acronyms the gender of which is not defined by the head of the acronym. In these cases, we have provided both the gender that the acronym head defines as well as the gender that is used commonly. For example, for the acronym *MAT*(see above) we provide both the feminine (*INDECfp*) and the neutral gender(*INDECnp*). In the forth field, the acronym analysis is provided, i.e. ΕΛΑΣ (ABBR=*Ελληνική Αστυνομία/National Police*), in order when NooJ encounters an acronym the acronym analysis to be given with the rest information of the acronym. With reference to loan acronyms, such as AIDS and NATO, that are used frequently in Greek language, we have lemmatized them like the Greek ones, and we have analyzed them both in Greek and English, i.e. AIDS+ABBR=*Acquired immune deficiency syndrome*/*Σύνδρομο Επίκτητης Ανοσολογικής Ανεπάρκειας* and NATO+ABBR=*North Atlantic Treaty Organization*/*Οργανισμός Βορειοατλαντικού Συμφώνου*.

4 Loan words written with Latin alphabet

In Greek texts not only loan acronyms written using the Latin alphabet can be met but words and expressions, too. This kind of lexemes and phrasemes occupy the third part of the present paper. They concern loans that have been introduced selfsame in Greek language and, thereafter, they have been lexicalized. Their lexicalization lies on the fact that they pose grammatical and semantic properties.

Totally 300 of such words and expressions of high frequency have been introduced into the macrostructure of the Greek NooJ dictionary. More particularly, we have included simple words (i.e. *eyeliner*), free phrasemes (i.e. *ladies and gentlemen, I love you, to be continued*), quasi-phrasemes (i.e. *hard disk, exit poll, super market, hot-dog*), full phrasemes (i.e. *What's up?, stand by*)as well as textual units (i.e. stereotypes, such as *c´est la vie, life is life*). It has to be mentioned that all these phrasemes are considered simple words by common Greek speakers, as they are not aware of the separate meaning of each phraseme´s element.

The microstructure of this dictionary is organized in six fields, following the lexicographic model of *monolingual coordinated dictionaries* of X. Blanco (2001), according to which we process the Greek NooJ dictionaries. The first two fields concern the grammatical and morphological properties of the lemmas. In the first part, the grammatical category of the entry is described (Noun, Adjective, Verb, etc) and, in the second, we assign the gender of the lemmas, just as it is used in the sentences of our corpus. For example, the neutral gender has been assigned to the word *look* (i.e. *το look του φετινού καλοκαιριού /the look of this summer*). Moreover orthographic variants of the lemmas in use in Greek language of the lemmas are provided, i.e. *hard disc/ hard disk* and *email*/*e-mail*.

The three last fields contain the semantic properties of the entries. Firstly, the syntactico-semantic features of the lemmas are given, following a seven-hold categorization (Human, Vegetable, Animal, Temporal, Locative, Concrete, and Abstract). Thereafter, the class of objects (i.e. emotion, professionals, clothes) of Gross (1992) and the domain (Mathieu-Colas & Buvet, 1999) (i.e. informatics, aesthetics, football) to which each noun belongs are annotated.

Thus, this dictionary, which has been incorporated to the Modern Greek NooJ dictionary, present the following structure:

email, e-mail,N+INDECn+Abst+Txt+Telecommunications

5 Conclusions

The purpose of this work was to enrich the database of the Modern Greek NooJ module. We focused on adverbs, acronyms and on loan words written in Latin alphabet, after having studied the linguistic analysis results´ of our corpora and having observed the necessity of the introduction of such units and their systematic description. Our work on adverbs will constitute a solid base for a thorough lexicographic treatment of Greek adverbs, taking as example the work of Català (2003), according to which the classes of verbs that are modified by each adverb will be annotated. Likewise, we will continue our study on acronyms and loan words written with Latin characters, in concrete, and we will, in general, keep on working with the existent lexicographic data and grammars as well as we will elaborate new ones in order to improve the automatic processing of the Greek language.

References

Anastasiadis-Simeonidis, A. 1986. *Neology in Standard Greek.* Thessaloniki: Aristotle University of Thessaloniki.

Baptista, J. 2004. Compositional vs. Frozen Sequences. *Journal of Applied Linguistics, Special Issue on Lexicon-Grammar* , pp. 81-92.

Blanco, X. 2001. Dictionnaires électroniques et traduction automatique espagnol-français. *Langages* (143), pp. 49-70.

Blanco, X., & Guenthner, F. 2004. Multi-lexemic Expressions: An Overview. *Lingvisticæ Investigationes. Revue internationale de linguistique française et générale. Hommages à Maurice Gross* .

Català, D. 2003. *Adverbes composés. Approches contrastives. Thèse de doctorat.* Barcelone: Univ. Autónoma Barcelona.

Foundation, M. T. 1998. *Dictionary of Modern Greek.* Thessaloniki: Manolis Triandafyllidis Foundation.

Gavriilidou, Z., Papadopoulou, E., & Chatzipapa, E. to appear. New data in the Greek NooJ module: A local grammar of proper nouns. In M. Silberztein, & T. Varadi, *Proceedings of the 2008 International Conference* (pp. 93-100). Cambridge Scholars Publishing.

Gavriilidou, Z., Papadopoulou, E., & Chatzipapa, E. 2008. The New Greek NooJ Module: morphosemantic issues. In X. Blanco, & M. Silberztein, *Proceedings of the 2007 NooJ International Conference* (pp. 96-103). Cambridge Scholars Publishing: Cambridge.

Gross, G. 1992. Forme d'un dictionnaire électronique. In *Actes du colloque La station de traduction de l'an 2000.*Mons.

Gross, G. 1996. *Les expressions figées : noms composés et d'autres locutions.* Paris: Ophrys.

Gross, M. 1986. *Grammaire transformationnelle.* Paris: ASSTRIL.

Mathieu-Colas, M., & Buvet, P.-A. 1999. Les champs Domaine et Sous-Domaine dans les dictionnaires électroniques. *Cahiers de Lexicologie* (75).

Mel'cuk, I. 1998. Collocations and Lexical Functions . In A. P. Cowie, *Phraseology: Theory, Analysis, and Applications* (pp. 23-53). Oxford: Clarendon Press.

Papadopoulou, E., & Gavriilidou, Z. 2010. Towards a Greek-Spanish NooJ module. In H. B. A., S. Mesfar, & M. Silberztein, *Finite State Language Engineering: NooJ 2009 International Conference and Workshop (Touzeur).* Centre de Publication Universitaire.

Silberztein, M. 2008. Complex Annotations with NooJ . In *Proceedings of the 2007 International NooJ Conference* (pp. 214-227). Newcastle : Cambridge Scholars Publishing.

Triantafyllidis, M. 1977. *Modern Greek Grammar.* Athens: ΟΕΔΒ.

Xydopoulos, G. J., & Vazou, E. 2007. Towards an account of acronyms/ initialisms in Greek. In M. Agathopoulou, E. Dimitrakopoulou., & D. Papadopoulou, *Selected Papers on Theoretical and Applied Linguistics* (pp. 231-243). Thessaloniki : Monochromia.

Assignment of Character and Action Types in Folk Tales

Piroska Lendvai[(1),] Tamás Váradi[(1)], Sándor Darányi[(2)], Thierry Declerck[(3)]

[(1)] *Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary,* {Piroska,Varadi}@nytud.hu

[(2)] *Swedish School of Library and Information Science, University College Boras/ Göteborg University, Sweden,* Sandor.Daranyi@hb.se

[(3)] *DFKI GmbH, Language Technology Lab, Saarbrücken, Germany* Declerck@dfki.de

Abstract

We process folk tales in order to extract two types of content descriptors coming from the Humanities domain: characters and their actions. We hypothesize that this will enable harvesting candidates for content units above word level. Our genre-specific corpus might be difficult to process by generic natural language processing tools, we thus develop NooJ lexicons and grammars for our purposes.

1 Introduction

Doubtlessly, the specifics of narratives reside well above word level, and possibly above phrase or sentence level as well. Narratology in the 20th century has come up with impressive theories with regard to the fundamental or minimal units of narratives, as experienced in different cultures, across genres and ages. One of these is actant theory which considers actants as behavioral patterns in a story situation, where one and the same actor can serve as different kinds of an actant depending on the situation (Greimas, 1966). Another relevant direction is the Proppian theory (Propp, 1968), which defines a limited set of prototypical functions (such as CONFLICT, KIDNAPPING, TRIAL OF HERO, etc.) building up the storyline, sequences of which in turn combine into a set of schemes of action plots. Within the various functions, prototypical characters perform characteristic roles, interacting with each other.

Propp stated that his goal was to "separate the component parts of fairy tales", in which "constants and variables are present", in order to determine "to what extent these functions actually represent recurrent constants of the tale"— for example, "Bába Jagá, Morózko, the bear, the forest spirit, and the mare's head test and reward the stepdaughter" (Propp, 1968). He states that "function is understood as an act of a character, defined from the point of view of its significance for the course of the action" (ibid.).

A number of computational models have been targeting the modeling and processing of fairy cf. e.g. (Dundes, 1965), (Maranda, 2010), (Lendvai et al., 2010) and their references.

In our previous work we noted that these largely lack the linking of linguistic objects with domain-specific elements of a given model, that are concepts and their relations from Humanities or Artificial Intelligence disciplines. It is however exactly the combination of domain-specific desriptors and linguistic markup that is a prerequisite for generating statistics of semantic and linguistic annotation to enable the finer-grained search and modeling that is still missing in Humanities research. For example, a folklorist might wish to find all verbs that express INTERDICTION in folk tales. The conceptual categories of Propp's classification scheme can thus serve as semantic markup elements of narrative events (well beyond the realm of folktale research), their representation is therefore important for improved retrieval in e-Humanities, in line with subgoals of several recent language technology projects such as CLARIN[1], D-SPIN[2], and AMICUS[3].

The elements of such an integrating annotation approach lend themselves to be implemented in NooJ, which is the focus of our current contribution. Soon after having started implementation, it became clear that the development of a general schema design would be necessary in order to accomodate the technical work in a solid framework that represents not only single elements, but their larger context as well. Based on our observations in the present paper, APftML, an augmented XML schema for fairy tales has been suggested by (Declerck et al., 2010), which is specified in detail in (Scheidel, 2010).

The current paper focuses on identifying folk tale character roles and their typical actions in such tales. By developing NooJ lexicons and grammars, we aim to semi-automatically collect invariants of these Proppian categories, annotating them in bootstrapped cycles. The importance of our text analytical approach is to regard actors in roles as slots that are filled in by values that are particular lexical items (pronouns, named entities, fairy tale entities), in order to generalize a tale's plot to a metalevel. Based on the approach and the markup scheme, in the long run we hope to be able to automatically capture motifs, i.e. semantico-structurally distinct and productive units in folk narrative texts. Propp's original goal with his work was to derive a morphological method of magic tale classification, based on the arrangements of functions. We hypothesize that particular value configurations of minimal role-action combinations will result in identifying specific subtypes of a tale.

2 Processing fairy tales

Our goal is to extract strings that typically occur in certain Proppian functions (e.g. *the river sheltered the little girl under its banks of pudding* lexicalizes a typical action of the function RESCUE OF HERO*),* as well as to obtain their transformations into patterns such as *HELPER sheltered HERO (under its banks of pudding),* where lexical items corresponding to character roles are mapped to a limited set of role labels. We use NooJ to experiment with extracting and labelling such domain-specific elements.

[1] http://www.clarin.eu/external/
[2] http://weblicht.sfs.uni-tuebingen.de/
[3] http://amicus.uvt.nl

One generally difficult issue in the realization of our goals is that we possess material that is unannotated: neither grammatical nor domain-specific information is assigned to the plain text files. Moreover, it is not trivial to establish the necessary processing pipeline either, since the underlying tasks depend on each other in a rather complex way. Both character and function detection depend on morphosyntactic analysis, which requires some adaptation of existing parsers to the specific data coming from the folklore domain. Even if one already knows morphosyntactic tags for the identification of character roles and actor-action patterns, they can only be labeled with certainty if one has identified the Proppian functions the patterns occur in — cf. (Scheidel, 2010) about the distribution of character roles over functions.

Our English corpus consists of tales from the Afanas'ev collection (Afanas'ev, 1945), while the Hungarian texts are coming from (Hermann, 2006)[4]; both are translations of the Russian originals. Below is an excerpt from "Nikita the Tanner" in English which can be considered a typical example of the corpus:

> *A dragon appeared near Kiev; he took heavy tribute from the people - a lovely maiden from every house, whom he then devoured. Finally, it was the fate of the tsar's daughter to go to the dragon. He seized her and dragged her to his lair but did not devour her, because she was a beauty. Instead, he took her to wife. Whenever he went out, he boarded up his house to prevent the princess from escaping.* (...)

2.1 Identifying characters

One of the goals of employing NooJ in the above context is to experiment with semantic markup, more specifically, to assign character roles to relevant string of words so that the above tale could be represented as

> *VILLAIN appeared near Kiev; VILLAIN took heavy tribute from the people, a lovely maiden from every house, whom VILLAIN then devoured. Finally, it was the fate of PRINCESS to go to VILLAIN. ...*

Propp established seven „dramatis personae", these are characters acting in certain roles during the course of the story plot:

- HERO: a character that seeks something;
- VILLAIN: opposes or actively blocks the hero's quest;
- PRINCESS: acts as the reward for the hero and the object of the villain's plots;
- her FATHER: acts to reward the hero for his effort;
- DISPATCHER: sends the hero on his quest;
- HELPER: aids the hero;
- DONOR: provides an object with magical properties;
- FALSEHERO: tries to take credit for the hero's deeds.

[4] Hungarian texts courtesy of Zoltán Hermann

By manually performing a pilot character role assignment procedure, we identified a number of difficult or controversial semantic representation and mapping issues in this task.

- Role transitions may take place during the course of a story, e.g. the *imp* (a sort of demon in another tale in the collection) now acts as a HELPER, at other times as DONOR, and could at certain points also be regarded as FALSEHERO;
- Some tales feature more than one instantiation of one and the same role, e.g. both the *Swan Geese* and *Baba Yaga* play the role of VILLAIN in the tale *The Magic Swan Geese*[5]; multiple persons may map to a single role: e.g. *tsar and tsarina* to FATHER, or a flock of birds (the *swan-geese*) to VILLAIN — i.e., roles might map to lexical elements that are morphosyntactically different from third person singular;
- Synonyms that stand for the same role occur in various distributions in a tale. For example, *the girl*, *daughter*, *maiden*, *the sister*, *she* may all map to HERO (*So the girl ran further, until she came to a river of milk flowing in banks of pudding*), but *she* in certain passages needs to be annotated as VILLAIN, because it refers to the witch (*When she had gotten the fire hot enough, she went to get the girl*);
- Another problem is the appearence of pronouns in direct speech as opposed to noun phrases in narrated passages. For example, when characters engage in dialogue, pronouns such as *I* and *you* may refer to any of the participating characters. *(I'm not going to eat your simple pudding with milk! I don't even eat the pudding we have at home.)*Such phenomena need to be treated based on dependencies of discourse analysis. So far our rule of thumb has been not to annotate pronouns that occur in direct speech.

The above attest that Propp's list of actors was generalized to abstract away from the individual text variant level, but this necessitates the design of a method for case-based decisions in order to obtain valid mappings between the text and character roles.

2.2 Identifying actions

The goal of this task is to harvest typical lexical items that characterize action, i.e. Proppian function types. There are 31 functions[6], each featuring several subtypes. We start out with a few assumptions, such as:

- Minimal Character + Action (+ Character) interactions take place on the sentence level;
- Typical interactions are expressed by NP+VP(+NP) sequences (*In swooped the swan-geese, snatched up the little boy, and flew away with him*);
- Certain lexical items are strong indicators of certain functions: e.g. INTERDICTION is often expressed by *don't + VP(Don't go out of the yard, …*).

[5] See http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/swan-geese.html
[6] See e.g. http://clover.slavic.pitt.edu/sam/propp/praxis/features.html

Such simple rules are able to extract a number of meaningful strings (resembling subject – verb – object triples) from each tale, cf. Fig. 1. Since the default English module in NooJ assigns only POS tags, the syntactically non-disambiguated material yields many irrelevant strings in this output.

Before	Seq.	After
A	dragon appeared near Kiev	; he took heavy tribute from
A dragon appeared near Kiev;	he took heavy	tribute from the people—a
to go to the dragon.	He seized her	and dragged her to his
her and dragged her to	his lair but	did not devour her, because
she was a beauty. Instead,	he took her	to wife. Whenever he went
took her to wife. Whenever	he went out	, he boarded up his house
wife. Whenever he went out,	he boarded up his house	to prevent the princess from

Figure 1. Triples resembling subject – verb – object combinations, extracted from *Nikita the Tanner*

2.3 APftML

The Augmented Proppian fairy tale Markup Language (APftML) is designed as a flexible multi-layer annotation scheme for fairy tales. By defining fine-grained schema elements, APftML enables alignment of the textual structure of a fairy tale text with several Proppian interpretation layers, accounting for improved representation of interdependent phenomena. APftML is an XML schema, a format which is supported in import/export operations of NooJ.

APftML brings two important changes in previous practices of converting Humanities descriptors into machine-readable (meta)data: it addresses annotation layers of both linguistic and narrative analysis, and also incorporates additional Proppian concepts which have not been considered for representation in machine-utilizable resources by previous approaches, e.g. character profiles, or attributes of a function.

3 Developments in NooJ

We address our goals by experimenting with NooJ, initiating modules holding the lexicon and local grammar of folk tale characters and actions. NooJ resources are created to enable identification of segment boundaries of the semantic descriptors we focus on.

3.1 Domain-specific lexical items

A domain-specific vocabulary holds items that are labelled as unknown by default lexical processing. These especially abound in Hungarian: we encounter archaic and dialectal word forms, lexical items denoting objects related to folklore and to fairy tales, as well as

semantically unusual compounds related to imagination. When listing these in a dictionary, two properties are added to them: *+Folk(lore)* by default, and *+Dial(ectal)* manually.

aludott, aluszik,<V+Past+Folk+Dial> ’slept’
aranycsengő, arany+csengő,<N+s+compound+Folk> ’golden bell’

+Dial items are linked to the lemma of which they are a terminological variant:

aluszik, alszik<V+Past+Folk+Dial> ’to sleep’
lóca, pad,<N+s+Folk> ’bench’

We can add small grammars of morphologically productive prefixes to capture domain-specific adjectives and nouns, e.g. ’golden-...’ (silver-, bronze-, iron-, diamond-, etc.):

arany+csengő, <N+s+compound+Folk> ’golden bell’
arany+lóca, <N+s+compound+Folk> ’golden bench
arany+nyereg, <N+s+compound+Folk> ’golden saddle’
arany+sörény+ű, <ADJ+compound+Folk> ’gold-maned’
arany+szőr+ű, <ADJ+compound+Folk> ’gold-furred’

The lexical constraint feature of NooJ can be used to check if components of such words are valid lexical entries in the —typically large-coverage and general— default English and Hungarian dictionaries.

3.2 Morphosyntactic properties in local grammars

The goal of local grammars is to provide guidelines for segmentation into Proppian functions and characters, producing dictionaries that enumerate folk tale characters, plus can additionally serve as gazetteers in lookup annotation procedures. We certainly need to capture higher syntactic complexity than can be covered by the simple assumptions set out in Section 2.2.

By experimenting with various simple syntactic grammars, a semantically meaningful pattern in folk tales turns out to be the infinitival clause (e.g. *<PART><V><PRO><N>**) that often expresses important goals (*to deliver her [from captivity], to free his land [from the wicked dragon]*) and conditions (*Upon receiving this letter*) of characters (cf. the concordance in Figure 2). Note in the left context pane of the concordance that the subject of these constructions is typically absent in this 5-word window, which indicates elliptic constructions — causing relatively high processing complexity even in these seemingly simple texts.

Before	Seq.	After
toward the dragon and began	to question him	. For a long time be
the Tanner in Kiev and	to send him	to deliver her from captivity
Kiev and to send him	to deliver her	from captivity. Upon receiving this
to deliver her from captivity.	Upon receiving this letter	, the tsar went in person
to beg Nikita the Tanner	to free his land	from the wicked dragon and
tsar in person had come	to see him	, he began to tremble with
little children and sent them	to implore him	, hoping that their tears would
devour him, then went forth	to give him battle	. Nikita came to the dragon

Figure 2. Infinitival constructions extracted from *Nikita the Tanner*

Grammars where subject pronouns are omitted will locate imperative phrases both in English and Hungarian, which mark certain functions, e.g. COMMAND (*Take care of your little brother!*). In Hungarian, imperative is somewhat easier to disambiguate, since it is expressed by a corresponding suffix sequence: the affix *(assimilated)j+<inflectional suffix>*. The Hungarian syntactic module of NooJ (Gábor, 2007) is able to detect predicates in imperative form (expressed by the same suffix as subjunctive mood), as well as NPs, so one can work along the lines of the pattern *<ADV>* <PRED+TenseMood=subj><NP>** and extract strings such as those in Figure 3.

Before	Seq.	After
néked ruhácskát, veszünk hozzá keszkenőt,	légy okos	, vigyázz a testvérkédre, az udvaron
veszünk hozzá keszkenőt, légy okos,	vigyázz a testvérkédre	, az udvaron túlra ne merészkedj
vigyázz a testvérkédre, az udvaron	túlra ne merészkedj	!" A szülők elmentek, a leány
míg egy kemencéhez nem ért. „	Mondd	, kemence, hová repültek a hattyú
vadalmafa akadt az útjába. „Vadalmafa,	mondd meg	kérlek, hová repültek a hattyúludak
kérlek, hová repültek a hattyúludak?" „	Egyél	a gyümölcsömből, akkor megtudod." „Ó
lekvárföveny! Merre repültek a hattyúludak?" „	Egyél belőlünk	, akkor elárulom!" „Ó, nálunk, apáméknál
A hattyú-ludak üldözőbe vették,	ha elkapják ezek a gonosz állatok	, nem lesz mit tenni! Odaért
most szemberepültek velük. Mi tévő	legyen	? Nagy a baj! Itt áll
vadalmafa. „Almafa-anyácska, rejts el!" „	Egyél az erdei almából	!" Gyorsan bekapta. Az almafa ágaival

Figure 3. Predicates in imperative extracted from the Hungarian translation of The Magic Swan-Geese (*Hattyúludak)*

3.3 Verbal valency

In Section 3.2 we saw that the simple grammar extracted meaningful phrases referring to actions, but often missed important arguments of the verb, such as in

to deliver her [from captivity]
to free his land [from the wicked dragon].

Here we extend the grammar to anticipate and cover such optional arguments of the verb phrase, representing them with NooJ's method for frozen expressions. We initiate a dictionary that holds a characteristic component of such items, associated with its syntactic properties and components, as described in (Silberztein, 2008). The dictionary links all components of such phrases together so that our folk tale texts can be annotated accordingly, identifying variations of strings in actual corpus data such as both *deliver*

her/the princess from captivity, or *free the princess/his land from the (wicked) dragon*. To abstract away from the actual values of arguments in these phrases, while storing their characteristic components, (pro)nouns need to be masked in the expressions. We will extend these representations into fully-fledged dictionaries as suggested by (Silberztein, 2008), e.g:

free, V+FXC+FLX=FREE+SHum+OFree+O2="from the <ADJ> dragon",*

where *OFree* specifies that the direct object of the verb *free* can take a value without constraints (e.g. *his land*, *the princess*, etc), whereas *O2* stores the frozen component (here between quotation marks) in a so-called property field.

3.4 Semantic categories

An elementary graph is made that describes typical conditions of identifying two characters, HERO and VILLAIN. Since the sheer context of nouns (NPs) denoting characters is too ambiguous to identify them (cf. the issues listed in Section 2.1), the grammar should cover complex conditions and their features. We aim to capture these via creating embedded graphs which represent particular Proppian functions typically performed by a given character (see Figure 4). These subgraphs hold strings that specify actions of only one or another character, taking these from the original text, (so far) without any abstraction. For example, the subgraph KIDNAPPING holds the string *seized her and dragged her to his lair*, whereas the subgraph STRUGGLE holds strings such as *killed* and *fought*. Thereby the grammar can partly solve ambiguous referencing, i.e. that *he* may refer to both HERO and VILLAIN. Note that the embedded graphs may not reflect the Proppian structuring of functions (cf. Section 2.2), where e.g. KIDNAPPING is established as subtype of VILLAINY — in further experiments we will examine whether the grammars can or should represent this. A presumably non-trivial technical issue in this respect is the recursive representation of characters in the grammars.

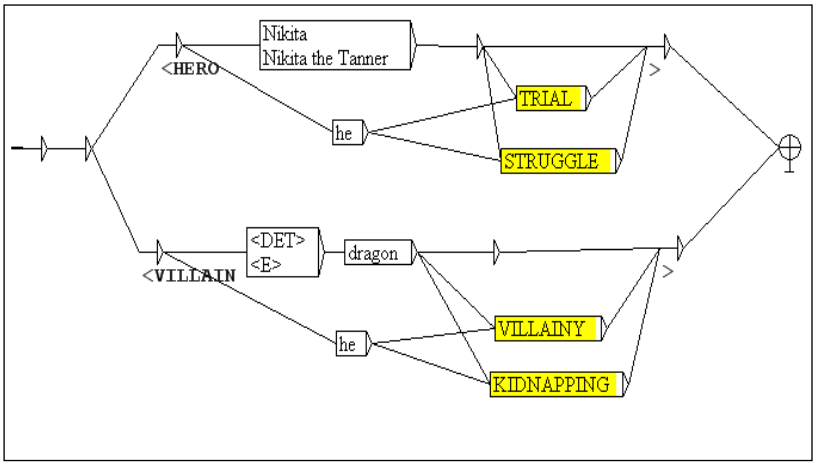


Figure 4. Initial graph for detecting HERO, VILLAIN, and some Proppian functions as embedded graphs (marked by yellow)

We anticipate difficulties related to the omission of grammatical subjects in the folk tale genre, as in the sentence below:

> ***He****seized her and dragged her to his lair but did not devour her, because she was a beauty.*
> *Elragadta* ***a sárkány****, elvitte a barlangjába, de nem falta föl, mert csodálta szépségét*

signaling the importance of resolving and assigning (character) tags on several segmentation levels next to actual strings found in the data. Identifying grammatical and semantic subjects is a prerequisite of correct attribution of characters, and requires co-reference and ellipse resolution on the clausal level, which is the target of ongoing work. Incremental processing of the texts (among others, by means of channeling the results of sections 3.2 and 3.3 into the subgraphs) is likewise under investigation. We hypothesize all the above to yield more fine-grained information on conditions and contexts for labeling Proppian descriptors.

4 Summary

Formalizing NLP tasks to facilitate Digital Humanities research is far from trivial since the underlying research issues are typically complex and vaguely formulated. We observed that folk tales syntax and semantics are dense and elliptic, featuring specific lexicons. Using NooJ, we attempted to map content desciptors to lexical variations, locating characters and actions. We showed that several means of representation of NooJ can be applied in the processing setup for the extraction and labeling of verbal chunks relating to character(s). Verbalizations of actions between a variety of folk tale characters can serve as gazetteers in lookup annotation procedures, and add a new layer of characterization of fairy tales in terms of their narrative semantics.

Modeling more tales from the Afanas'ev corpus will result in richer and more consistent output. Generalizing the results will shed light on the computational applicability of Propp's method not only for cultural heritage corpora, but in other narrative genres as well.

References

Afanas'ev, A. 1945. "Russian fairy tales". Pantheon Books: New York.

Declerck, T., A. Scheidel, Lendvai P. 2010. "Proppian Content Descriptors in an Augmented Annotation Scheme for Fairy Tales". In: *Proc. of the International Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.*

Dundes, A. 1965. "On Computers and Folktales". *Western Folklore* 24:185-189.

Gábor, K. 2007. “Syntactic Parsing and Named Entity Recognition for Hungarian with Intex”. In: *Formaliser les langues avec l’ordinateur: De Intex à NooJ*. Besançon: Presses Universitaires de Franche-Comté.

Greimas, A.J. 1966. “Sémantique structurale”. [Structural semantics.] Larousse, Paris.

Hermann, Z. (ed), 2006. “A tűzmadár. Varázsmesék A. Ny. Afanaszjev mesegyűjteményéből”. [The Firebird. Magic tales from the collection of A.N. Afanas’ev.] Magvető, Hungary.

Lendvai, P., T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, F. Peinado 2010. “Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case”. In: *Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC’10).*

Maranda, P. 2010. “Morphology and Morphogenesis of Folktales and Myths”. In: *Proc. of the 1st International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Comminication Texts*.

Propp, V.J. 1968. “Morphology of the folktale”. University of Texas Press: Austin.

Scheidel, A. 2010. “An Augmented Annotation Scheme for Fairy Tales”. Bachelor’s Thesis, Saarland University.

Silberztein, M. 2008. “Complex Annotations with NooJ”. In: *Proc. of the 2007 International NooJ Conference.*

Named Entities in Chinese

Ying Yang[(1)], Miloš Utvić[(1)], Gordana Pavlović-Lažetić[(1)], Duško Vitas[(1)]
[(1)]University of Belgrade, Serbia

Abstract

Our research aimed at Recognition of Named Entities in Simplified Chinese text and the improvement of alignment methods, based on named entities. We experimented with the Englishand Chinese versions of the novel Around the World in 80 Days, written by Jules Verne. Using ACIDE, which integrated XAlign and Concordancer tools, we aligned the English and Chinesetext to produce bitext. We then built linguistic resources for NooJ, such as dictionaries to help locate named entities, and graphs to deal with more complicated patterns.

1 Introduction

Information Extraction (IE) is a type of information retrieval (IR) technology that automatically maps natural-language text into structured relational data, i.e. categorized and contextually and semantically well-defined data.
The significance of Information Extraction has been unveiled, as the amount of information available in unstructured form is experiencing exponential growth. The Internet is a case inpoint. Through information extraction, knowledge can be made more accessible by means of transformation into relational data, or by marking-up with XML tags.
The term **Named Entity (NE)**, was first introduced in the Message Understanding Conferences (MUC), it's now a widely used term in Information Extraction (IE), Question Answering (QA) and other Natural Language Processing (NLP) applications. On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person,organization, and location names were marked as well as dates, times, percentages, and monetary amounts, as described in Chinchor (2003). **Named Entity Recognition (NER)** (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories of named entities.
According to Vitas and Krstev (2006), a **bitext** is a merged document composed of two versionsof a given text, usually in two different languages. An aligned bitext is produced by analignment tool or aligner, which automatically tries to synchronise (align or match) the different versions of the same text, considering some level of text structure (chapters, paragraphs,sentences). Named Entities can be used as anchors to guide the alignment process, as describedin Bonhomme and Romary (2000).

2 Basic Introduction to Chinese

Chinese is comprised of Pinyin (phonetics) and Hanzi (Chinese characters). Pinyin is the Romanization system for Chinese. It literally means "spelling sound". For instance, the word 塞尔维亚(Serbia) has Pinyin writing (sài ěr wéi yà), each part corresponding to one character.Chinese characters are transcribed into Roman alphabets to help provide a visual

representation of Chinese sounds. Nowadays Pinyin is also used as one common typing method to enter Chinese characters into computers and cell phones.
There are tens of thousands of Chinese characters. The Chinese dictionary *Zhonghua cihai* which was published in the year 1994, collected 87019 characters. However, it is estimated that basic Chinese literacy can be achieved with knowledge of 2,000 to 3,500 characters. The Chinese characters are logographic symbols. Each individual character represents an idea or thing. The combinations of characters express different meanings, which are usually but not necessarily the combinations of each character's meaning.

2.1 Two Standard Sets of Written Chinese

Currently there are two standard sets of Chinese characters of the contemporary Chinese written language: **Simplified Chinese** and **Traditional Chinese**. Simplified character forms were created by decreasing the number of strokes and simplifying the forms of a sizable proportion of traditional Chinese characters. It was promoted to increase literacy. Most of the simplified Chinese characters in use today are the result of the works moderated by the Chinese government in the 1950s and 60s. And nowadays, simplified Chinese is used in mainland China, Singapore, Malaysia and traditional Chinese is used in Taiwan, Hong Kong, Macau and some overseas Chinese communities.
The only big difference between the Traditional and Simplified Chinese is that some oftheir corresponding characters are written differently. The following are some examples, the left side are Traditional Chinese characters and the right side are their simplified forms:
對→ 对 *(dui, correct)*
觀→ 观 *(guan, view)*
風→ 风 *(feng, wind)*
潔→ 洁 *(jie, clean)*
鄰→ 邻 *(ling, neighbor)*
極→ 极 *(ji, extreme)*
廣→ 广 *(guang, vast)*
寧→ 宁 *(ning, quiet)*

Traditional Chinese should not be confused with the **ancient Chinese**. The traditional Chinese only refers to the writing of the characters, while the ancient Chinese is a concept of awhole language system including writing and syntax, which is very different from the contemporary Chinese we use. It's hard even for a Chinese individual to comprehend ancient Chinese, if he/she has never learnt it.

2.2 Characteristics of Chinese

Chinese syntax is, in a way, similar to English. Sentences are often formed by stating a subject which is followed by a predicate. The predicate can be an intransitive verb, a transitiveverb followed by a direct object, a linking verb followed by a predicate nominative, etc. The most common sentence structure has SVO (subject + verb+ object) word order.
Some of its characteristics that are relevant to our project are:
• Chinese does not have tenses. Tenses are indicated by adverbs of time ('tomorrow',

‘just now’) or particles.
• Chinese does not use grammatical gender.
• There is no grammatical distinction between singular or plural, the distinction is accomplishedby sentence structure.
• All words have only one grammatical form. No changes in the form of the wordthrough inflection of verbs according to tense, mood and aspect.
• Chinese sentences are written as character strings with no spaces between words.

2.3 Chinese NE Recognition Issues

There are two major problems hampering Chinese NE recognition.
The first is the segmentation of words. In Chinese, there are no spaces to delimit the words. Sometimes even native speakers couldn’t agree on the *right* segmentation. Take the segmentationof words in this sentence for example:

世界上国际象棋、围棋和中国象棋三大棋类中…
(Eng. *Among the top three most popular board games throughout the world, chess, weichiand Chinese chess…*).

There can be more than 2 ways of segmentation.

世界| 上| 国际| 象棋| 、| 围棋 |和| 中国| 象棋| 三| 大| 棋类| 中 |…
and

世界上| 国际象棋| 、|围棋| 和| 中国象棋| 三大| 棋类| 中 |…

As word segmentation is the basic initial step to almost all linguistic analysis tasks, many techniques developed in English NLP cannot be applied to Chinese.
Second, there is no exterior feature (such as the capitalization) to help identify the Named Entities.
As noticed in Ye *et al* (2002), word is a vague concept in Chinese, being defined as consisting of one or more characters representing a linguistic token. Words in Chinese are actually not well marked in sentences, and there does not exist a commonly accepted Chinese lexicon Zhang *et al* (2000).

3 Bitext Alignment

In our project, we took the Chinese and English translation texts from the famous JulesVerne's novel *Around the World in 80 Days* for exploration.

In general, the bitext construction proceeds in two main steps:

- Segmentation of text into sentences.
- The alignment of the sentences.

3.1 Segmentation of Text into Sentences

According to Vitas and Krstev (2006), the common methods of alignment of a bitext usually assume that before alignment both texts have been marked up, which means that the elements of its logical layout were explicitly and unambiguously annotated. Extensible Markup Language (XML) is used to tag the logical layouts. The marked-up XML document can be viewed as a tree structure that has leaf nodes and labeled internal nodes.
In the Jules Verne's novel *Around The World In 80 Days*, we tagged every chapter with aheading, and a main text which is divided into paragraphs and segments, as illustrated in Table1.

<body> <head>第一章斐利亚·福克和路路通建立主仆关系</head> <seg> 1872 年，白林敦花园坊赛微乐街七号（西锐登在1814 年就死在这听住宅里），住着一位斐利亚·福克先生，这位福克先生似乎从来不做什么显以引人注目的事，可是他仍然是伦敦改良俱乐部里最特别、最引人注意的一个会员。</seg> … … <seg>福克先生就只是改良俱乐部的会员，瞧，和盘托出，仅此而已。</seg><seg>如果有人以为象福克这样古怪的人，居然也能参加象改良俱乐部这样光荣的团体，因而感到惊讶的话，人们就会告诉他：福克是经巴林氏兄弟的介绍才被接纳入会的。</seg><seg>他在巴林兄弟银行存了一笔款子，因而获得了信誉，因为他的账面上永远有存款，他开的支票照例总是"凭票即付"。</seg> … … <seg>现在赛微乐街的寓所里只剩下路路通一个人了。</seg> </body>	<body> <head>Chapter I IN WHICH PHILEAS FOGG AND PASSEPARTOUT ACCEPT EACH OTHER, THE ONE AS MASTER, THE OTHER AS MAN</head> <seg>Mr. Phileas Fogg lived, in 1872, at No. 7, Saville Row, Burlington Gardens, the house in which Sheridan died in 1814. He was one of the most noticeable members of the Reform Club, though he seemed always to avoid attracting attention;</seg> … … <seg>Phileas Fogg was a member of the Reform, and that was all.</seg><seg>The way in which he got admission to this exclusive club was simple enough. He was recommended by the Barings, with whom he had an open credit.</seg><seg> His cheques were regularly paid at sight from his account current, which was always flush.</seg> … … <seg> Passepartout remained alone in the house in Saville Row.</seg> </body>

Table 1 The Segmentation of the Bitext of a Chapter from the Novel Around the World in 80 Days

The methods of segmentation are applied to each of the two texts separately. The units are usually sentences, but they can also be larger, as paragraphs, or smaller, as words. Interestingly, one of the familiar circularities of computational linguistics, namely the fact that sentence shave to be marked before processing, though that processing itself will determine what the sentences are, is present in the alignment problem as well. Once sentences are tagged, segment alignment could be applied.

3.2 The Alignment of the Sentences

Tagged texts can be processed by alignment systems, for instance by tools XAlign and Concordancier, developed within Loria (2006) and based on statistical methods. The goal of the alignment is to establish 1:1 relations on the segment level.

In our project, we used ACIDE system (Aligned Corpora Integrated Development Environment), described in Obradović, Stanković and Utvić (2007). It integrated Loria alignment tools (XAlign and Concordancier) and tools for exporting to TMX (Translation Memory eXchangeformat) and HTML format of aligned texts.

The input texts to XAlign (as shown in Figure 1) must be XML files. The 37 chapters from the novel *80 Days Around The World*, were all segmented in the previous phase. The MultiAlignProperties file specifies structure tags and the way they'll be treated by Loria tools.

After specifying the paths to input texts in XAlign and path to output directory, and activation of the button 'Align', the output window will show the message about the result of XAlign work. If everything runs without an error, ACIDE will create 3 files: *aligned_f_id.xml*,*aligned_s_id.xml*, *aligned_fs.xml*, and the last one, containing alignment links, will be openedin Concordancier, as shown in Figure 2.

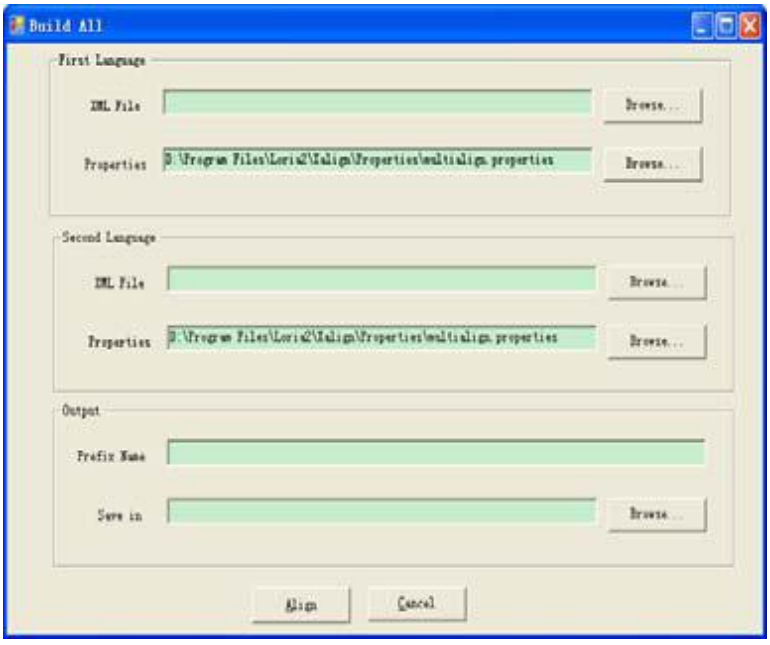


Figure 1. XAlign

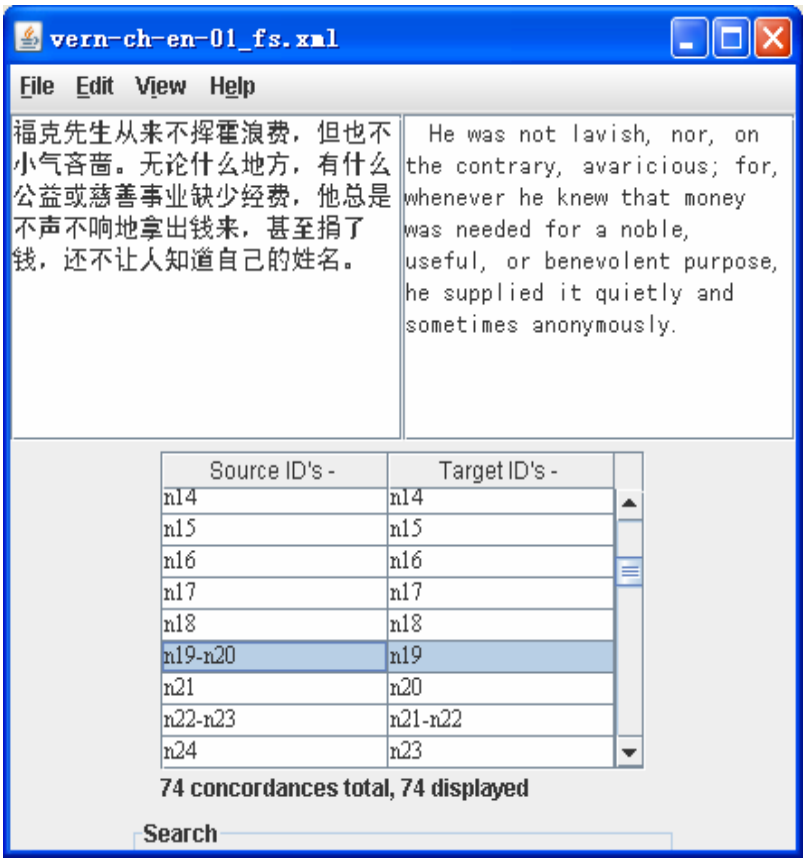


Figure 2. Concordancier

However the bitext is not always well matched in XAlign, as there are some shortcomings in statistic models. Using lengths of sentences as indications of correspondence in the bitext space may work well among western language due to their substantial similarities, but when it comes to East-Asian and Western language bitext, which have little in common, this method is not as efficient. Therefore a Concordancier comes in handy to rematch the segments when there is discordance.

User can click on the numbered segments to 'unlink' and 'link' the segments manually inorder to align them correctly. And by clicking on Source ID's column, user can sort translation units, and on particular ID to view the specified translation unit.

4 Named Entities Recognition in NooJ

NooJ provides several ways for pattern locating, and the ones that we used here are "regular expressions" and "grammar graphs", as described in Silberztein (2010).

4.1 Dictionary Building

The Chinese module that we downloaded from the NooJ website[1] was based on Traditional Chinese (produced by Huei-Chi Lin, Université de Franche-Comté), which is mainly used in Taiwan, Hong Kong and some overseas Chinese communities and not applicable to the Simplified Chinese text. In order to process Simplified Chinese, we built its own dictionaries first.
We made dictionaries including countries, currencies, numbers, measurement units, month, day, verbs and so on. The following are some examples of the dictionary entries:

英里,*N+Measurement+Length#mile*
分钟,*N+Measurement+Time#Minute*
星期三,*N+Day#Monday*
英镑,*N+Currency#Pound*
伦敦,*N+City#London*
做,*V#do*
块,*Q#a piece of*

As we can see, the Chinese dictionaries are relatively simpler, for there is no infection, plural or singular form, grammatical gender or tense.
But as mentioned in the section of the characteristics of Chinese, word is a vague concept. The combination of Chinese characters can yield to innumerable words, phrases, whose numberare growing even larger as the time goes on. And there is no commonly acknowledged lexicon to be applied. Making decant Chinese dictionaries is no easy work.
To make the exploitation of the novel *80 Days around the World* easier, we also made simple dictionaries of cities, names, etc, that are based on the novel.
After compiling the dictionaries and building the morphological grammar graphs, we could use them to explore the text.

4.2 Pattern Searching using Regular Expression

Take names for example.
Type <*N+Name*>in the regular expression pattern locating window in NooJ, and search. We can find patterns in the result (Figure 3).

[1] http://www.nooj4nlp.net

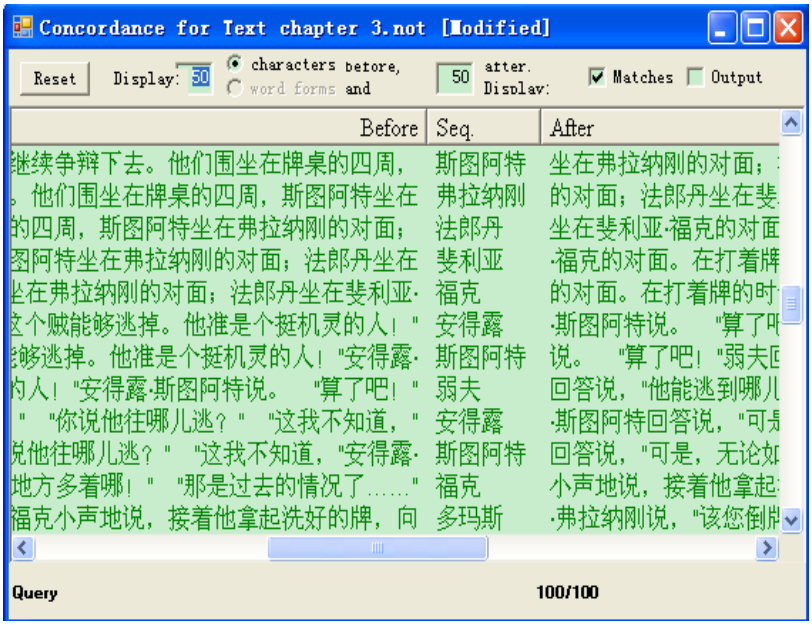


Figure 3. NooJ Concordances (<***N+Name***>)

They are 斯图阿特(Stuart), 弗拉纳刚(Flanagan), 法郎丹(Fallentin), 斐利亚(Philea), 福克(Fogg), 安德露(Andrew), etc. Indeed, the results returned are the names in the text. And to locate only the last names in the text, just repeat the procedure, and type in <*N+Last+Name*>in locate configuration. And as shown in Figure 4, first names such as 斐利亚(Philea), 安德露(Andrew), are eliminated.

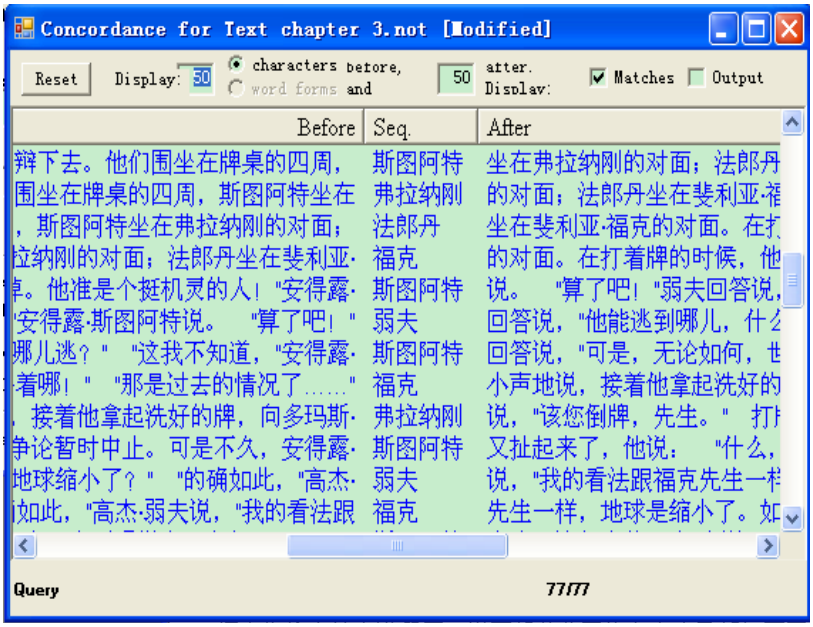


Figure 4 NooJ Concordances (<***N+Last+Name***>)

4.3 Pattern Locating Using Grammar Graphs with Time Expression Example

Regular Expression pattern locating is easy and convenient, however it becomes cumbersome when the expression is complicated. In such situation, graphs would be a more desirable way to describe.

Taking the expression of time in modern Chinese for example, the following is the illustration of time expression recognition from Chinese text using NooJ.

In Chinese there are several ways to express time. For every full hour, they are all expressed in the same form:

time on the hour + 点/时*[o'clock]* + 正*[sharp](optional)*.

For instance, 12:00 can be called as 十二*[twelve]*点*[o'clock]* or 十二*[twelve]*点*[o'clock]*正*[sharp]*, or using Arabic numerals, it would be 12 点(正). This is the same for other time on the hour, from 0-24. To express that pattern, a graph in the Figure 5 would be appropriate.

However the pattern *Number* +点/时 doesn't just mean special time, it can appear in the scoring of games as well, such as poker. It is because 点, while carrying the meaning of *o'clock*, also means *point(s)*. Apart from this, in Chinese 一点 can mean *one o'clock*, or *a little*, and the latter is a very common case. The system can't differentiate between these two. In a word, this pattern searching sometimes may present more than we are looking for.

Every hour on the hour is just one special case of 'time'. When the minute hand of a clockis not pointing to the big '12' on the clock, we have to come up with other ways in which time is likely to be told.

Examples:

12:05
十二*[twelve]*点*[o'clock]*（零*[zero]*）五*[five]*分*[minute]*, and this '零*[zero]*' is optional.
十二*[twelve]*点*[o'clock]*过*[past]*五*[five]*分*[minute]*, five minutes past twelve.
12:10
十二*[twelve]*点*[o'clock]*十*[ten]*分*[minute]*, twelve ten.
十二*[twelve]*点*[o'clock]*过*[past]*十*[ten]*分*[minute]*, ten minutes past twelve.
12:15
十二*[twelve]*点*[o'clock]*（过*[past]*）十五*[fifteen]*分*[minute]*, twelve fifteen, fifteen minutespast twelve.
十二*[twelve]*点*[o'clock]*一*[one]*刻*[quarter]*, a quarter past twelve.
12:30
十二*[twelve]*点*[o'clock]* （过*[past]*）三十*[thirty]*分*[minute]*, twelve thirty, thirty minutespast twelve.
十二*[twelve]*点*[o'clock]*半*[half]*, half past twelve.
While 过means *past*, which is often used in the first 30 minutes within an hour, for theminutes past 30, we use 差, which means *lack*.
12:45
十二*[twelve]*点*[o'clock]*四十五*[forty-five]*分*[minute]*, twelve forty-five.
十二*[twelve]*点*[o'clock]*三*[three]*刻*[quarter]*, three quarters past twelve (very rarely used).
一*[one]*点*[o'clock]*差*[lack]*一*[one]*刻*[quarter]*, a quarter to one.
一*[one]*点*[o'clock]*差*[lack]*十五*[fifteen]*分*[minute]*, fifteen minutes to one.

This almost wraps up all the possibilities in modern Chinese of the way in which time istold. And to indicate "a.m." and "p.m." of the day, a list of words used to describe the time inthe day is added, for instance, 早上(morning), 傍晚(at dusk), 晚上(evening), 深夜(late night).
The ancient Chinese time expression is a totally different concept, and it is rarely used orseen except in the ancient literature. Unfortunately we have to bypass that to avoid the confusion.
The following is what the graph, which contains all the time expression patterns, looks like.

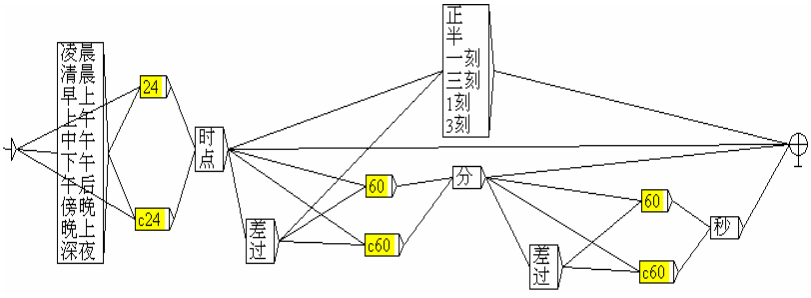


Figure 5. Graph for the Time Expression

4.4 Pattern Locating Using Grammar Graphs with Measurement Example

In the dictionary *measurement.dic*. we listed all kinds of measurements: length, time, area,volume, pressure, weight, temperature, etc.

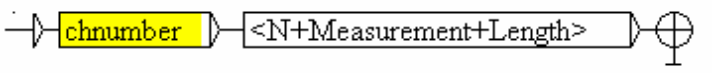


Figure 6. Measurement

Figure 6 is an example of how to find all the length expressions in the text which has a specific number. <*N+Measurement+Length*>in the graph finds all the length entries as<*N+Measurement+Weight*>finds all the weight entries in the dictionary. That is how we defined them in the dictionaries earlier. *Chnumber* is a subgraph describing numbers in Chinese.
Apply this graph to the text.

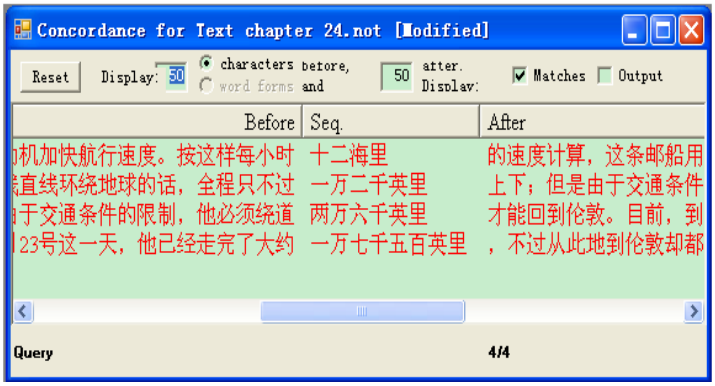


Figure 7. Length Measurement Search Result

The results are: 十二海里 (twelve nautical miles), 一万两千英里(twelve thousandmiles), 两万六千海里(twenty-six thousand nautical miles),一万七千五百英里 (seventeen miles).
And if we combine a time measurement and a length measurement in a way shown in the Figure 8, we can get something new, i.e., speed. The meaning of both Chinese characters 每and 一is “every”. The grammar graph (Figure 8) is equivalent to *every* ⊕*a unit of time* ⊕*number* ⊕*a unit of length* which is a usual way to describe speed, for example, 每分钟二十米 (every minute twenty meters).

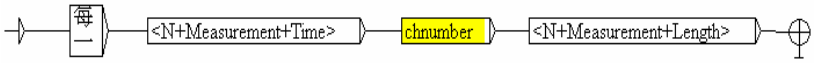


Figure 8. Speed Expression Using Measurement Dictionary

Search results for grammar graph (Figure 8) are shown in Figure 9.

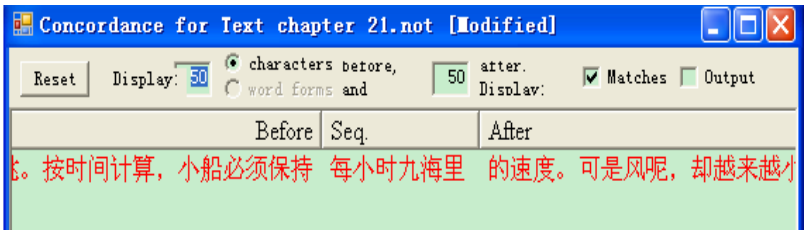


Figure 9. Speed Measurement Search Result

The result 每小时九海里 means “nine nautical miles each hour”, a perfect pattern of speed in the text.

5 Conclusion

The results obtained achieved our purpose. However, it was limited because the lexical resources we collected and built to analyze the Chinese text with were not comprehensive. If there is a further study or research on it, the establishment of Simplified Chinese module, and the refinementand enrichment of the linguistic resources might be the next consideration.

Keynotes:
e-dictionaries, named entities, information extraction, Simplified Chinese, NooJ

Acknowledgments

This work has been financed by the Ministry of Science and Technological Development, proj. no. 148921A.

References

Bonhomme, P. & Romary, L. 2000. “Parallel alignment of structured documents”, In J.Véronis (ed), *Parallel Text Processing*, 233-253.

Chinchor , N. A. 2003. “OVERVIEW OF MUC-7/MET-2”.
http://www.nlp.org.cn/docs/20030724/resource/overview_of_MUC7.pdf

Loria. 2006. XAlign and Concordancier (Alignement multilingue)
http://led.loria.fr/en_outils.php#4.
http://led.loria.fr/en_outils.php#5

Obradović, I. & Stanković, R. & Utvić, M. 2007. “Integrisano okruženje za pripremu paralelizovanog korpusa”, In B. Tošović (ed), *Zbornik radova međunarodnog simpozijumaRazlike između bosanskog/bošnjačkog, hrvatskog i srpskog jezika*, Graz, Austria,563-578. (English translation: “Aligned Corpora Integrated Development Environment”, In B. Tošović (ed), *Proceedings of the International Conference Differences among Bosnian, Croatian and Serbian Language*).

Silberztein, M. 2010. *NooJ manual* http://www.nooj4nlp.net/NooJManual.pdf

Vitas, D. & Krstev, C. 2006. “Literature and Aligned Texts”, In M. Slavcheva & G. Angelova & K. Simov (eds) *Readings in Multilinguality*, Institute for Parallel Processing, BulgarianAcademy of Sciences, Sofia, Bulgaria, 148-155.

Ye, S. & Chua, T. & Jimin, L. 2002. “An Agent-based Approach to Chinese Named Entity Recognition”, In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1. P1 – 7*. International Conference On Computational Linguistics.

Zhang, J. & Gao, J. & Zhou, M. 2000. “Extraction of Chinese Compound Words - An Experimental Study on a Very Large Corpus”.
http://www.aclweb.org/anthology/W/W00/W00-1219.pdf

Recognition of Libyan Person names using NooJ platform

Abdelsalam Almarimi[(1)], Abdelmajid Ben Hamadou[(2)], Khaled Hussain[(1)], H. Fehri[(2)]

[(1)]College of Electronic Technolgy, Baniwalid, Libya
[(2)]Research Laboratory Miracl, ISIM-Sfax, Sfax University,Tunisia

Abstract

The objective of this paper is to propose a system for the recognition of Libyan Person Names in order to translate them into English using the NooJ platform. In fact, Libyan Person Names have syntax quite specific and are based on a particular vocabulary.
This article begins by presenting the specificity of the Libyan Person Names on the level of the vocabulary and the local syntax. After that, we detail the recognition approach based on syntactic rules. Then, we present the implementation of the recognition approach using the platform NooJ. The evaluation of the developed system is performed on a test corpus created from Libyan newspapers, Specific web sites and Scholar Lists of student names from institutions covering the main Libyan regions.
The developed system will be improved by the translation component. It will be made available to academic institutions, municipalities and governments to help them to recognize and translate properly this type of Named Entities.

1 Introduction

Named Entities Recognition (NER) has become the last few years an interesting research domain. It is very useful for many applications such as web mining and information retrieval. Several works has been performedespecially for Latin languages and mainly for English (Daille 2000). Very modest works has been dedicated to Arabic NER (Mesfar 2007), (Mesfar 2008), (Fehri et al. 2008), (Fehri et al. 2009), (Shaalan et al. 2009).

This paper focuses on the recognition of Libyan Person Names (LPN) which has specific syntax and vocabulary. We propose a rule based system for the recognition of Libyan Person Names. The implementation of this system is performed using the NooJ platform [Silberstein 2005]. The rest of the paper is organized as follows. We begin by detailing with the specificity of the Libyan Person Names and the associated recognition problems. Then, we give a list of main patterns related to the LPN. Finally, we present the implementation of the proposed system using Nooj Platform (Silberstein 2005). This implementation is followed by the result of the evaluation of the system using a learning corpus for testing.

Specificity of the Libyan Person Names

Libyan Person Names are usually composed of two parts: the first name and the last name. The last Name is composed of one name. The first name of a can be simple like "Abdel Kharim عبدالكريم" or composed of two names like "Mohamed Abdel Kharim محمد عبدالكريم". It can be followed by the name of the father and grand father in order to remove ambiguities between persons especially in the official documents.

Person Names can occur in text individually (First Name and Last Name) or preceded by triggers word(s), which must be defined by the article *the*: ال. The trigger represents a honorific title like "الأستاذ*Professor*" or a professional title like "المدير*Director"*. Itcan be simple,composed of one word as "الصيدلانية*druggist*" or complex: a combination of honorific titles, professional titles and adjectives like:
"الأمين العام*The General Secretary"*.

Also, LPN are not always completely written in the text (i.e., truncated), especially for well known persons or when a name occurs more than once.
قدم **د. علي عبيد** محاضرة قيمة... ثم أعطى **د. علي** جملة من الاقتراحات.
In the second part of the sentence the Last Name is elided**"د. علي"***Dr. Ali"* , in state of **د. علي عبيد"***Dr. Ali Obaid*".

2 Recognition problem

The recognition of Libyan Person Names is facing classical NER problems, problems due to the specificity of the Arabic language (on the orthographic level and the absence of standards when writing Named Entities) and specific problems due to cultural aspects. Those problems can be summarized in the following points:

- The problem of the right boundary (Left boundary for Arabic):

Difficulty of recognizing the end of a Person Name because of the variability of the number and the constitution of the different Person Name components: Trigger First Name and the parent link. Also, some components of the Person Name are voluntary elided (truncated):

  - Triggers can help to recognize the beginning of a PN, but they are not always composed of one word. They have variable length and their own local grammar.
  - The length of the First Name is variable.
  - The truncated Person Names and the necessity to use the context.

- The problem of localizing the beginning of the Person Name when the trigger is absent. This problem comes from the fact that First Names can be common nouns or adjectives, like "صالح Salah" which is the Arabic equivalent of the adjective *useful*.

- The existence of triggers can help to recognize the beginning of a PN, but they can pose some problems related to the fact that we can find the name of an organization between the trigger and the Person Name, like:

  مدير إدارة التفتيش وحماية المستهلكأ. الهاشمي ضيف الله الفقهي

  Director of the Administration of control and consumer protection Prof. Al-Hachmi thif Allah Al-Feki

  مدير: *Director*is the trigger

  أ. الهاشمي ضيف الله الفقهي **:** *Prof. Al-Hachmi thif Allah Al-Feki*

is the Person Name
إدارة التفتيش وحماية المستهلك: *the Administration of control and consumer protection* is the name of an organization.

- Ambiguïty between the First Name and the last Name. The First Name can also be Last Name and vice versa like :
محمد عمران*Mohamed Omran* and عمران التاجوري*Omran Al-Tajouri*.
Omran is the Last Name in the first example and the first Name in the second example.

- complexity of the Arabic morphological and spelling system:
  - Agglutination can lead to ambiguity like : باسم : Preposition ب +Noun اسم and باسم*Basem* is a Fist-Name
  - Spelling : confusion between different forms of the same grapheme (Archi-grapheme) :
    - confusion between ي*Yaa* and ى *alif* such as: the name علي is written على.
    - confusion between أ*Hamzah* and ا *alif* such as: the name أكرم is written اكرم.
- Punctuations are agglutinated to the Person Names or triggers such as الدكتور/*Doctor*.

3 Libyan Names Patterns

Libyan Person Names Patterns was extracted semi-automatically from a learning corpus that we constructed for this purpose. We generate useful concordances by using the Nooj regular forms facility.
We discover two kinds of Libyan Person Names Patterns: explicit patterns and contextual patterns. Explicit patterns correspond to complete LPN and contextual patterns correspond to incomplete LPN.

We distinguish four explicit patterns (Pattern1 to Pattern4) and three implicit patterns (Pattern 5 to Pattern 7) for the Libyan Person Names:

<Pattern 1> := <First-Name><Last-Name>
عمر المختار*Omar Al-Mokhtar*

<Pattern 2> := <First-Name><Parents-link><Last-Name>
علي بنصالحالتويجري*Ali bin Salah Al-Towaijri*

<Pattern 3> := <Defined-Trigger><First-Name><Last-Name>
المهندس علي التويجري*Engineer Ali Al-Towaijri*

<Patern 4> := < Defined-Trigger><First-Name><Parents-link><Last-Name>
المهندس علي بنصالحالتويجري Engineer Ali bin Salah Al-Towaijri

<Patern 5> := < Defined-Trigger><First-Name>
المهندس علي Engineer Ali

<Patern 6> := < Defined-Trigger><Last-Name>
المهندس التويجري Engineer Al-Towaijri

<Patern 7> := <Abbreviate-Trigger><First-Name><Last-Name>
أ.د. علي المختار *Prof. Dr. Ali Al-Mokhtar*

<Patern 8> := <First-Name><Parents-link><Last-Name>
د. علي بنصالحالتويجري Dr. *Ali bin Salah Al-Towaijri*

4 Learning corpus

The learning corpus is important component in our approach. It is used to discover the different patterns of the Libyan Person Names and to elaborate the following dictionaries:

- Dictionary of the First Names composed of more than 4000 entries
- Dictionary of the Last names composed of about 2000 entries,
- Dictionary of the component of triggers composed of more than 100 entries.

This corpus was collected from:

- Journalistic articles dealing with different topics (Politics, Economy, Medicine, sports, culture, education, etc.) from the following Libyan newspapers:
  - Libya Al-Yaom
  - Al-Shams
  - Al-Fajjr Al-Jadeed

  The length of the collected articles is 211520 words (see figure 1)

- Specific web sites

- Scholar Lists of student names from institutions covering the main Libyan regions

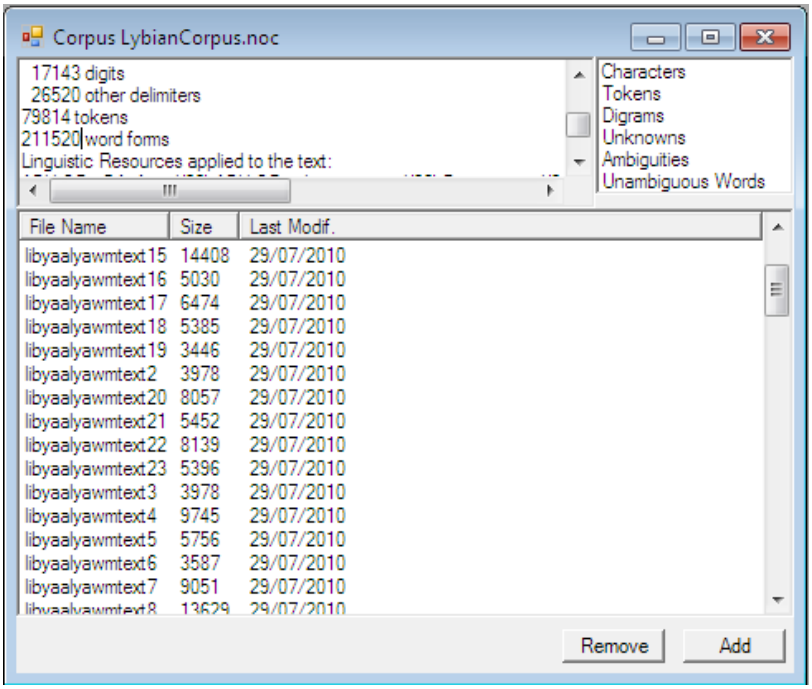


Figure 1. Journalistic corpus in HTML format.

5 Nooj implementation

5.1 Dictionaries:

To recognise the Libyan Person Names, we create three main dictionaries:

- The dictionary of the Libyan First-Names composed of entries. Each entry contains the following information: First Name, Grammatical information (N + FIRST-NAME) and the translation into English.
- The dictionary of the Libyan Last-Names composed of entries. Each entry contains the following information: The Last Name, Grammatical information (N + LAST-NAME) and the translation into English.
- The dictionary of the different components of the triggers. In addition to the trigger and the grammatical information, each entry contains the inflectional paradigm in order to recognise different generated forms (feminine such as الدكتورة, plural, dual, …).

5.2 Recognition graphs

We built the corresponding recognition graphs, considering different patterns as following:

- The Main Graph corresponding to the recognition the Libyan Person Names

- The Title Graph corresponding to the recognition of the trigger
- The FirstName graph recognising the First name with the eventual Parent link
- The LastName Graph recognising the Last Name

Figure 2 shows the Main Graph with a conjunction of the indicated graphs. We see in particular that the Libyan Person Names can begin directly with the first Name, First Name or the Last Name can be elided.

Main Graph

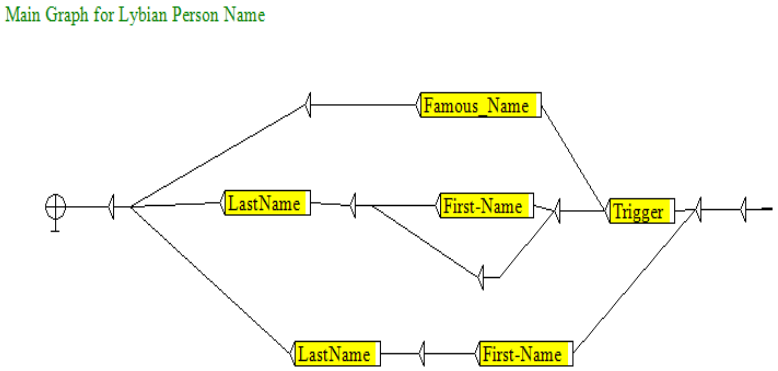


Figure 2. The main graph for LPN recognition

Figure 3 shows the complexity of the trigger structure.

Title Graph

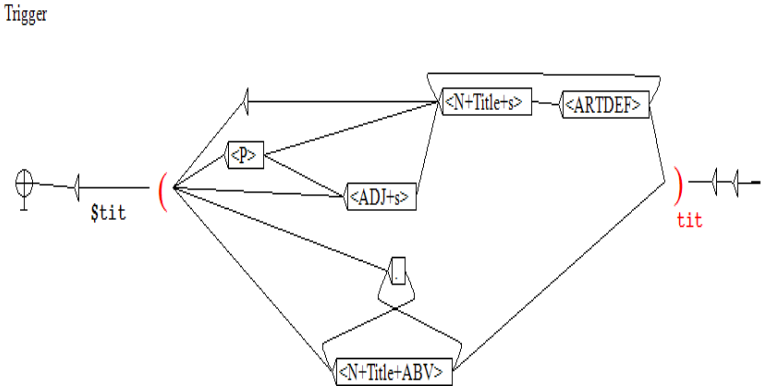


Figure 3. Libyan Person Names Title/Trigger structure

5.3 Concordency

The execution of the proposed system on the test corpus gives the following concordances. We can see that an ambiguity is generated when both the First Name and the Last Name are common nouns or adjectives. This ambiguity can not be removed without using the context. It is the case of the sequence: انتصار جديد*New victory* and the sequence على السرير*On the bed***.**

We can also see that the system do not recognize Person Names that are not Libyan. It is the case of the Person Name محمد عامر*Mohamed Ameur* (cf. ficure 4, line 10).

Figure 4 gives en extract of the concordances.

After	Seq.	Before
بتحقق لصالح الجانب الليبي. وحول	انتصار جديد	معمر القذافي) الجمعة الماضية أنه)
أمين لجنة إدارة سوق الأوراق (	سليمان الشحومي	ليبيا فاق توقعاتي 10/01/2010 وقع
نائب رئيس مجلس إدارة البورصة (	د. محمد عمران	) السنج . والإنفاقية الثانية كانت مج
ليبيا ستستضيف مؤتمر اتحاد البورصات	للدكتور محمد عمران	فنيا 100% . ليبيا اليوم : سؤال أخير
عضو هيئة التدريس بكلية الفنون (	الدكتور (محمد صبري الصالحي	هذه الفنون لدى المختصين. وألقى
أمين اللجنة الشعبية العامة (رئيس (	البغدادي المحمودي	) المدينة بعد الانتهاء منها. وكان
بمحاضرة تناولت فيها تعريف هذا (	للدكتورة (خيرية السائحي	بهذا الفيروس، حيث كانت البداية
أخصائي الأمراض المعدية بالمركز الوطني (	الدكتور (محمد قمو	وطرق الوقاية منه. كما ألقى
والدكتور (محمد عامر) بمحاضرة علمية (	الدكتور (محمد الهنشيري	احصائيات منظمة الصحة العالمية. وألقى
عامر) بمحاضرة علمية مشتركة حول	والدكتور (محمد	(العالمية. وألقى الدكتور (محمد الهنشيري
01/2010 لا يزال السجين السابق (حسن	حسن القذافي القماطي	) تفاقم الحالة المرضية للسجين السابق
يعاني آلاما شديدة بعد تفاقم (	حسن القذافي أحمد محمود القماطي	10/01/2010 لا يزال السجين السابق
بمكبلا بأغلال من الحديد، ، لم	علي السرير	تقلي إلى المستشفى مقيدا وأنا
رقم الحساب 32489 بمصرف الوحدة فرع	حسن القذافي	العون والمساعدة لهذا الشاب المكلوم
هذا الإنجاز العلمي أحد المشاريع (	الدكتور (رمضان قشوط	رئيس قسم هندسة الطيران بالكلية
رئيس المنظمة الوطنية للشباب الليبي (	الهادي الحويج	وتقوية القرارات العربية. (عبد
وزير الخارجية الليبي السابق ورئيس	علي عبد السلام التريكي	دبلوماسي بين ليبيا وسويسرا. وأدرج
للصحفيين في مقر بعثة البلاد	عبد الرحمن عبد الرحمن شلقم	منسوب ليبيا لدى الأمم المتحدة
أن رئيس الجمعية العامة "يؤكد	باسم التريكي	تتح سوى الآن. وقال المتحدث
رئيس تحرير جريدة أخبار ليبيا (	عاشور الشامس	) خميس) قائد بارز بالجيش. وقال
لأن الغرب لم يعطه المزيد	أمل القذافي	موقف ليبيا الحازم الجديد حيبة
للصحافيين أن عرابيس دقت برن (	عبد الرحمن شلقم	) عويل بين البلدين. وقال السفير
تصوير / النعاس خميس بإشراف وتنظيم	النابر البعقوبي	/ الحديثة في الجوسسة المتقدمة متابعة
أيها الأخوة والأخوات الحضور..باسم .	الدكتور / عبد السلام المريمي	أولها كلمة اللجنة التحضيرية . ألقاها
والدكتور / عبد السلام المريمي والدكتور	الدكتور خالد التكراط	وعلى حسن الضيافة وأخص بالذكر
والدكتور علي حسين الذي استقبلنا	عبد السلام المريمي	/ بالذكر الدكتور خالد التكراط والدكتور
الذي استقبلنا ورافقنا طوال السفر	علي حسين	والدكتور / عبد السلام المريمي والدكتور
لجميع المشاركين في هذه الورشة	الأستاذ الدكتور محمد منصور الشريف	الهيئة الوطنية للبحث العلمي وأمينها
ورقة بعنوان / التعرف على أنواع	للدكتور علي حسين	بينها: محاضرة بعنوان / الحوسبة السحابية
ورقة بعنوان / تقنيات إخفاء المعلومات	محمد كعجمان	كأساس لبناء نظام فقهي خبيرللاستاذ
و فرج الغزالي. في اجتماع	عبد الله المبروك	حشيم وعبد الرحيم إدريس و
في اجتماع المجلس التنفيذي والجمعية	فرج الغزالي	و عبد الله المبروك و
للشمس: جامعة ناصر الأممية لها	الدكتور محمد المدني	للأحداث بوجدان عربي متأثر للوصل
الذي عاصر النادي طويلاً ويشرك	الزميل يوسف الشيباني	حوى معلومات دقيقة وقيمة.. ولاننسى

Figure 4. The concordance obtained by the execution of the system on a test corpus.

6 Evaluation of the system

In order to evaluate the proposed LPN recognition system, we built a test corpus composed of the learning corpus and new students' lists and Wikipedia pages.

The evaluation metrics we used are Recall, Precision and F_measure (2*P*R/(P+R)). Let's remember that the recall measures the quantity of relevant responses of the system compared to the ideal number of responses; Precision is the number of relevant responses of the system among all the responses he gave and the F-measure is a combination of

Precision and Recall for penalizing the very large inequalities between these two measures. The values obtained in the evaluation of our work are:

	Precision	Recall	F-measure
Libyan Person Names	89%	77%	82.5%

The lack of precision is related to the fact that Proper Names can be common nouns such as the recognized sequence: انتصار جديد which is not a Person Name and means *new victory*. Lot of PN are not identified in the learning corpus. Most of them are not Libyan Person Names.

7 Conclusion and Perspectives

In this paper, we presented a Libyan Person Name recognition System based on a set of rules and three dictionaries. Rules are implemented as transducers in Nooj platform. Some ambiguities require the use of context. A second pass considering the right context is necessary to remove them.
The developed system was evaluated on a test corpus. The obtained results are very promising.
As perspectives, we will continue to improve the system performance by resolving some ambiguities and extending dictionaries. We also plan to integrate a translation/transliteration component for Libyan Person Names and develop, on the basis of the proposed system, social applications for municipalities and Passport administration to provide right translations for Libyan Person Names.

References

Ben Hamadou, A., Piton, O., Fehri, H. 2009. *Recognition and translation Arabic-French of Named Entities: case of the Sport places*. NOOJ'09, Tozeur 2009.

Fehri H., Haddar, K. Ben Hamadou, A. 2009. *Translation and Transliteration of Arabic Named Entities.* LTC'2009, Polland 6-8 November 2009.

Mesfar, S. 2007. *Named Entity Recognition for Arabic Using Syntactic grammars.* NLDB 2007 Paris, 28-38.

Mesfar, S. 2008. *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard*. Thèse, Nov. 2008, Université de Franche Comté.

Poibeau, T. and Kosseim, L. 2001. *Proper Name Extraction from Non-Journalistic Texts.* Proc. Computational Linguistics in the Netherland 2001.

Shaalan, KH. and Raza, H. 2009. NERA: Named Entity Recognition for Arabic. *Journal of the American Society for Information Science and Technology*, Vol. 60, N° 8, pp: 1652-1663, August 2009.

Silberztein, M. 2005. NooJ's dictionaries. *Actes de la conférence Internationale LTC*, 2005, Poznan, Pologne.

Acknowledgement. This work was supported by the **National Association of Scientific Research**, The Higher Education, **Tripoli, Libya.**

Improved Parser for Simple Croatian Sentences

Kristina Vučković[(1)], Božo Bekavac[(1)], Zdravko Dovedan Han[(1)]

[(1)]*Department of Linguistics,*
Faculty of Humanities and Social Sciences,
University of Zagreb
Zagreb, Croatia.

Abstract

In this paper, we will present the work that has been done to improve the existing syntactic parser presented at the NooJ 2009 conference. We will show and explain the grammar for detecting nominal predicate in a simple sentence. The nominal predicate in Croatian language is made of the auxiliary verb 'to be' and an <NP> in Nominative case. The <NP> can be a complex <NP> made of a single noun and any number of adjectives, pronouns and numbers proceeding that noun and agreeing with it in number, gender and case, but also a single noun, a single pronoun, a single adjective or even an adverb. A problem of coordination of two or more <NP> nodes of different gender and its agreement with the main verb in the cases where coordination is a subject of a sentence will be discussed. The work will further enlighten and discuss other important properties of Croatian sentence complexity. At the end of the paper, the results will be evaluated through precision, recall and f-measure to show the adequacy of the model.

1 Introduction

This work is done inside a framework of building a parser for Croatian (Vučković, 2009, Vučković et al., 2009). The paper will present the FSTs or syntactic grammars for recognizing and annotating nominal predicate and coordination of two or more <NP> nodes (Silberztein, 2008) of different gender when that <NP> is playing the subject role in a sentence.

Our goal is to come as close as possible to perfect syntactic disambiguation of all Atomic Linguistic Units (ALUs) in the sentence (Silberztein, 2009). So far we are still working on simple sentences, i.e. sentences with only one <VP> chunk whether it is a continuous or a discontinuous chunk. However, the grammars for recognizing complex sentences are just being developed (see Štefanec *et al*. in this volume, Vučković *et al*, 2010).

2 Nominal Predicate

The nominal predicate in Croatian language is made of the auxiliary verb 'to be' and an <NP> in nominative case as shown in Figure 1. In some cases, the <NP> could be in instrument or genitive case but these occurrences will not be discussed in more details in this paper (Barić, 2005).

The <NP> (Vučković *et al*, 2008; Vučković, 2009; Vučković *et al*, 2010) can be a complex <NP> made of a single noun and any number of adjectives, pronouns and numbers

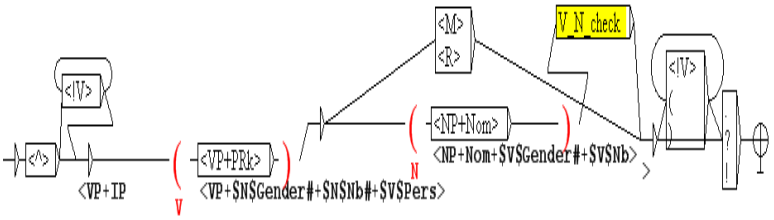


Figure 1.

proceeding that noun and agreeing with it in number, gender and case, but also a simple <NP> made of a single noun, a single pronoun, a single adjective, a single number or even an adverb.

*On je **dječak**.* (He is **a boy**.)

*On je **moj**.* (He is **mine**.)

*On je **mlad**.* (He is **young**.)

*On je **prvi**.* (He is **the first**.)

*On je **tamo**.* (He is **there**.)

*On je **moj mladi prijatelj**.* (He is **my young friend**.)

If the nominal predicate is made of an auxiliary verb and an <NP>, single noun, single pronoun or single adjective, than the verb and the nominal part have to agree in gender and number (Figure 2).

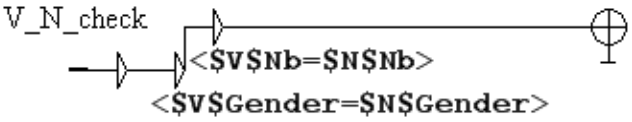


Figure 2. Checking the agreement betweenthe verb 'to be' and its nominal predicate

Of course, there are exceptions that do not comply to these rules, like:

*A.) Još uvijek **smo traumatizirano društvo**.*

- Still we **are the traumatized society**.

*B.) Bizovačke toplice **su oaza** slavonskog turizma.*

- Bizovacka Thermal Springs **are an oasis** of Slavonian tourism.

In the example **A**, since the subject is in plural form, the auxiliary verb from the nominal predicate is also in plural, but the nominal part is in singular.

In the example **B**, semantically, the <NP> *'Bizovačke toplice'* is in singular since there are only one such thermal springs but there is no singular form for the word '*toplice*' so the VP that follows this word as a subject of the sentence has to follow it in the plural form as well. However, the nominal part of that VP is in singular '*oaza*'. Exceptions like these are not described with the grammar in Figure 1 and will need some further attention in our future work that will probably include addition of some new annotations on the dictionary level.

If there is an agreement between the verb '*to be*' and the nominal predicate, they are both disambiguated on a syntactic level so that only the ALU's of matching Gender and Number of <VP> and <NP> part remain in the TAS. Furthermore, the <VP> and <NP> obtain a new joint ALU **<VP+IP>** which indicates the nominal predicate chunk.

3 Coordination of multiple <NP> nodes

A problem of coordination of two or more <NP> nodes of different gender and its agreement with the main verb in the cases where coordination is a subject of a sentence will be discussed (see Figures 3, 4, 5, 6 and 7).

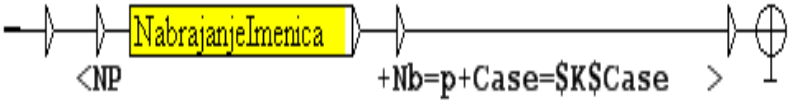


Figure 3. Main graph for <NP> coordination

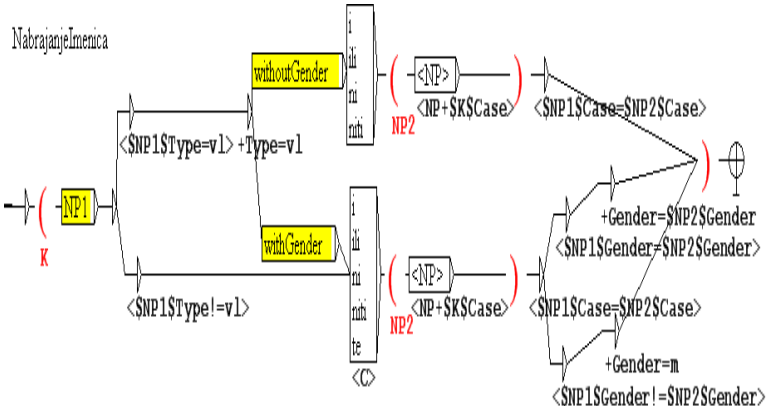


Figure 4. Level 1 Subgraph for <NP> coordination

Coordination of two or more <NP> chunks can be observed through the following four groups of coordination:

1. All <NP>s are of the same gender => coordination = gender of the nouns (see Figures 3, 4, 5 and 7)
    - **<NP+f**<$_{NP+f}$ *jabuka*>, <$_{NP+f}$ *kruška*> *i*<$_{NP+f}$ *šljiva*>**>**
        - an apple, a pear and a plum
2. All <NP>s are in feminine and at least one is in masculine gender => coordination = masculine gender (see Figures 3, 4, 5 and 7)
    - **<NP+m**<$_{NP+f}$ *jabuka*> *i*<$_{NP+m}$ *ananas*>**>**
        - an apple and a pineapple
3. All <NP>s are in feminine and at least one is in neutral gender => coordination = masculine gender (see Figures 3, 4, 5 and 7)
    - **<NP+m**<$_{NP+f}$ *jabuka*> *i*<$_{NP+n}$ *slovo*>**>**
        - an apple and a letter
4. All <NP>s are in neutral and at least one is in masculine gender => coordination = masculine gender (see Figures 3, 4, 5 and 7)
    - **<NP+m**<$_{NP+n}$ *slovo*>*i*<$_{NP+m}$ *ananas*>**>**
        - a letter and a pineapple

Coordination of two or more proper nouns like geographical names or names of people (except where the last names are only given) follows the same concept as previous <NP>s (see Figures 3, 4, 5 and 7):

Geographical names:

masculine and masculine => masculine

- **<NP+m**<$_{NP+m}$ Zagreb> i <$_{NP+m}$ Dubrovnik>**>**

masculine and feminine => masculine

- **<NP+m**<$_{NP+m}$ Zagreb> i <$_{NP+f}$ Barcelona>**>**

Names of people:

masculine and masculine => masculine

- **<NP+m**<$_{NP+m}$ Tin Ujdur> i <$_{NP+m}$ Filip Kocijan>**>**

feminine and masculine => masculine

- **<NP+m**<$_{NP+f}$ Ema Ujdur> i <$_{NP+m}$ Filip Kocijan>**>**

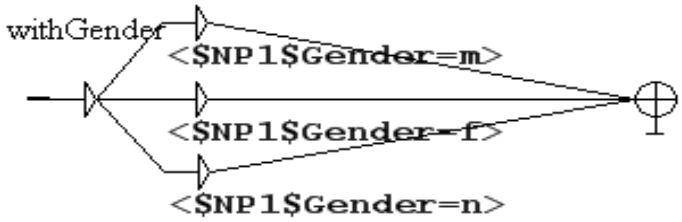


Figure 5. Level 2 Subgraph - checking the Gender of an <NP>

Coordination of two or more last names only is a challenge since last names do not have a gender (see Figures 3, 4, 6 and 7):
<NP+m<$_{NP+m}$ Ujdur> i <$_{NP+m}$ Kocijan>> su otišli…

- *Ujdur and Kocijan left …*

<NP+f<$_{NP+f}$ Ujdur> i <$_{NP+f}$ Kocijan>> su otišle…

- *Ujdur and Kocijan left …*

<NP+m<$_{NP+f}$ Ujdur> i <$_{NP+m}$ Kocijan>> su otišli…

- *Ujdur and Kocijan left …*

However, the gender of the genderless coordination i.e. the 'UNDEFINED' gender (see Figure 6), may be inferred from the verb form since it depends on the gender, but unfortunately, not for all verb tenses (Vučković, 2009).

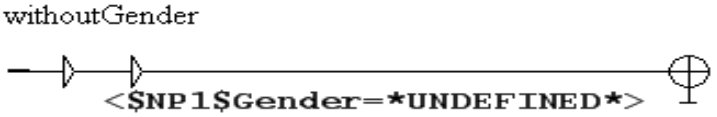


Figure 6. Level 2 Subgraph - checking if the gender is undefined

The grammar also disambiguates each <NP> involved in the making of coordination (see Figures 4 and 7) so that only the ALUs of matching case attribute remain in the TAS. The coordination <NP> is further marked with a shared ALU as a plural <NP> with the matching case **<NP+Nb=p+Case=KCase>** (see Figure 3) and gender defined according to the rules of previously defined groups of coordination (see Figure 4).

Figure 7. Level 2 Subgraph for recognizingany number of <NP> nodes followed by a comma

4 Results and Discussion

We used the Croatia Weekly 100kw (CW100) corpus (cf. Tadić 2002, Vučković et al. 2008) to extract a small gold standard for purposes of this experiment. We chose two sets of simple sentences. The first one consisted of 150 sentences that were used for the development stage and another one consisted of 155 simple sentences that were used for the purposes of evaluation. Sentences from both sets were randomly chosen considering that they consisted of only one verb phrase, whether dislocated or not, and any number of noun phrases, prepositional phrases, conjunctions, adverbs, numerals, exclamations and/or particles.

The Table 1 shows the performance of the system in terms of precision, recall and F1-measures for the recognition of nominal predicates, <NP> coordination but also for the recognition of all sentence parts in general.

	Sentences	Nominal predicate	<NP> Coordination
Precision	0,660	1	0,958
Recall	0,980	1	1
F1-measure	0,789	1	0,978

Table 1. Measures for recognition of Sentences, Nominal predicate and <NP> coordination

From Table 1 we learn that the system, although it performs perfectly for the recognition of nominal predicates, its performance is somewhat decreased in the case of the <NP> coordination recognition and sentence parts recognition in general. Let us elaborate.

All the occurrences of the <NP> coordination not annotated correctly are due to the incorrect tagger, meaning that they were wrongly marked as a part of an <NP>. Such is the following example xml annotated (see sentence [C]) where underlined chunk is incorrectly marked as an <NP+Nom> i.e. word '***dalje'*** is tagged as an adjective instead as an adverb since both have the same form:

[C] ***Oni i dalje mirno*** *gledaju u svoje užasne poslove i njihovu povijesnu katastrofu.*
(They are still quietly looking at their terrible jobs and their historic catastrophe.)

<SENTENCE>
<SUBJECT> ***Oni i dalje mirno*** **(They are still quietly)** </SUBJECT>
<PREDICATE> *gledaju* (looking)
</PREDICATE>
<PREPOSITIONAL PHRASE>*u svoje užasne poslove i njihovu povijesnu katastrofu*
(at their terrible jobs and their historic catastrophe)
</PREPOSITIONAL PHRASE>.
</SENTENCE>

Poor precision for the sentence parts recognition can be explained with ambiguous annotations of some sentence parts. Some <PP>s are thus marked both as **indirect object** and as **prepositions of time, place or manner**, and some <NP>s are marked both as **subject** and **direct object** of the sentence. Such are the following examples:

[D] ***Taman veo*** *pokrio je* ***logor***. (**Dark veil** covered the **camp**.)

<SENTENCE TYPE="Sub_Pred">
<SUBJECT>
<OBJEKT TYPE="DIREKTNI">*Taman veo* (Dark veil) </OBJEKT>
</SUBJECT>
<PREDICATE> *pokrio je*(covered) </PREDICATE>
<OBJEKT TYPE="DIREKTNI">
<SUBJECT> ***logor*** **(camp) </SUBJECT>**
</OBJEKT>.
</SENTENCE>

Sentence [D] has two chunks with ambiguous annotations. The subject of a sentence <*Taman veo*> is also marked as a direct object while the direct object <*logor*> is marked both as a subject and as an object of a sentence.

[E] *Aron je sjedio* ***pored rijeke***. (Aron was sitting **by the river**.)

<SENTENCE TYPE="Sub_Pred">
<SUBJECT> *Aron* (Aron)
</SUBJECT>
<PREDICATE> *je sjedio* (was sitting) </PREDICATE>
<OBJEKT TYPE="Indirekt">
<PREPOSITIONAL PHRASE>*pored rijeke* (by the river)
</PREPOSITIONAL PHRASE>
</OBJEKT>.
</SENTENCE>

In sentence [E] the chunk <*pored rijeke*> is ambiguously recognized as an indirect object and prepositional phrase although it should be only marked as a prepositional phrase of place.

To solve these ambiguities we will need to add some additional semantic information to our lexicon and also expand existing syntactic grammars in order to eliminate subject/object and object/prepositional phrase ambiguities.

5 Conclusion

In order to obtain perfect syntactic disambiguation of a Croatian text, our attention was given to two very important language occurrences: the nominal predicate and the problem of coordination of <NP>'s of a different gender. Both instances are quite common and frequent in texts making their solvent necessary and important at this early stage of parsing Croatian texts. Although some instances still remain unsolved or ambiguous, we believe that Croatian partial parser is well on its way of becoming a full parser.

Key words

Croatian, parser, simple sentences, nominal predicate, coordination, syntactic grammars, NooJ.

Acknowledgments

This work was done within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grant 130-1300646-1776 and 130-1300646-1002.

References

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević V. and Znika, M. 2005. *Hrvatska gramatika*, Školska knjiga, Zagreb.

Silberztein, M. 2003. *NooJ Manual*, available at the web site http://nooj4nlp.net (200 pages).

Silberztein, M. 2008. "Complex Annotations with NooJ". In X. Blanco, M. Silberztein (eds) *Proceedings of the 2007 International NooJ Conference*. Cambridge Scholars Publishing, Barcelona, 214-227.

Silberztein, M. 2009. "Syntactic parsing with NooJ". In A. Ben Hamadou, S. Mesfar, M. Silberztein (eds) *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*, Centre de Publication Universitaire, 177-189.

Silberztein, M. 2010. "Disambiguation Tools for NooJ". In T. Varadi, J. Kuti, M. Silberztein (eds) *Applications of Finite-State Language Processing – Selected Papers from the 2008 International NooJ Conference*. Cambridge Scholars Publishing, Budapest.

Tadić, M. 2002. "Building the Croatian National Corpus". In M. Gonzalez Rodriguez, C.P. Suarez Araujo (eds) *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC2002*. ELRA, Paris-Las Palmas, 441-446.

Vučković, V. 2009. *Model parsera za hrvatski jezik*, PhD dissertation, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Vučković, V., Bekavac, B., Dovedan, Z. 2009. "SynCro - Parsing simple Croatian sentences". In A. Ben Hamadou, S. Mesfar, M. Silberztein (eds) *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*, Centre de Publication Universitaire, 207-217.

Vučković, K., Tadić, M., Dovedan, Z. 2008. "Rule Based Chunker for Croatian". In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, D. Tapias (eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC'08*, Marrakech, ELRA, 2544-2549.

Vučković, K., Tadić, M., Dovedan, Bekavac, B. 2010. "Croatian Language Resources for NooJ". In V. Lužar-Stiffler, I. Jarec, Z. Bekić (eds) *Proceedings of the 32nd International Conference on Information Technology Interfaces*, SRCE University Computer Centre, University of Zagreb, Zagreb, 121-126.

Vučković, K., Agić, Ž., Tadić, M. 2010. "Sentence Classification and Clause Detection for Croatian". In M. Tadić, M. Dimitrova-Vulchanova, S. Koeva (eds) *Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages*, Croatian Language Technologies Society, Faculty of Humanities and Social Sciences, Zagreb, 131-138.

Towards Parsing Croatian Complex Sentences: Dependent Noun Clauses

Vanja Štefanec, Kristina Vučković, Zdravko Dovedan Han

Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia

Abstract

In this paper, authors will present a model for partial parsing Croatian complex sentences in which a dependent clause serves as a direct object to the predicate in the main clause. This research is based on the resources that have already been developed for parsing simple Croatian sentences (Vučković et al., 2010b).
So far, sentences that we were able to parse using these resources are of the basic structure consisting of a subject, predicate, direct and indirect object, and adverbial of time and place. Model we shall present in this paper will extend this structure to the following sentence structure <main clause <object clause>>. Our primary indicator for this type of sentence will be the absence of the required direct object in the main clause as well as the presence of one of the subordinating conjunctions ('da', 'kako') or complementizers (relative pronoun, adverb of place, time, cause or manner) which usually introduce the object clause in Croatian.
Since this type of complex sentences is very common, we chose it to test the adequacy of this method for its potential use in describing other types of dependent clauses in Croatian language. At the end of the paper, we will evaluate the adequacy of the model through precision, recall and F-measure.

1 Introduction

Building a rule-based parser for a loose syntax language like Croatian presents quite a challenge. Although one might think that analyzing syntax of a language that has most of its syntactical relations "hidden" in morphology is a rather trivial problem, this is definitely not the case with Croatian. Within a framework of the Croatian module for NooJ the parser for Croatian is being built with new improvements constantly made to it. For now, our parser can analyze Croatian simple sentences with high accuracy (Vučković *et al*., 2010b). In this paper, we decided to take things one step further towards parsing complex sentences by identifying the dependent clause within complex sentence. For the purpose of this experiment we have chosen to describe probably the most frequent of all dependent clauses in Croatian – the object clause.

2 Our motivation

To be able to go into parsing of dependent clauses, the first thing we have to do is to find a way to determine their boundaries within the complex sentence. Basically, there is a need for some kind of clause-splitting pre-parsing method for identifying series of chunks which

are bound with strong syntactical connections, i.e. clauses. And that's exactly what we'll be dealing with in this paper.
The benefit of performing this analysis as a pre-parsing method is twofold; firstly, we're limiting the number of possible annotations by focusing the parser on the parts of the sentence that have to be independently analyzed, and secondly, since this analysis highly depends on the output of the chunker, we can perform disambiguation of chunks to some extent, as well as identify the most frequent chunker mistakes and work on the improvements.

Although not a new method, the clause splitting or clause identification is rarely written about in the field of natural language processing. Ejerhed (1988) has compared rule-based and stochastic methods for finding clauses in unrestricted text for the purpose of detecting large prosodic units in text-to-speech system. Leffa (1998) has developed a rule-based method for clause processing in the English/Portuguese machine translation system. Leffa's method is especially interesting because it reduces the whole clause to one word (noun, adjective or adverb) and by doing that transforms complex sentence into simple. Orasan (2000) and Ram and Devi (2008) both have investigated hybrid methods for clause identification in which linguistic rules were used for improving the results. Some foundations for text segmentation in Croatian can be found in Boras (1998).

In this work we have focused only on object type of clauses, but similar approach can be applied for identification of other dependent clauses that could simplify and improve the parsing process (Vučković *et al.*, 2010a).

3 Overview of the work

We have composed a local grammar that will recognize the dependent noun clause behaving as a direct object to its superordinate-clause predicate. This type of the dependent noun clause will be referred to as the object clause. The grammar can recognize two syntactic constructions behaving as an object: simple object clause and coordination of any number of object clauses.

This is done simply by defining the co-text in which this kind of clause can occur without going into description of its structure. The reason for this will be explained later (see Section 4). In defining the preceding and succeeding co-text we are relying mostly on the output of the chunker but simpler syntactical elements like individual morphological categories, as well as punctuations, are also taken into account.

For the annotation of object clauses we used general <CLAUSE> tag with three attributes: Type, Subtype and Sense. In the present case, value of the attribute Type will be "obj", and values of the other two will denote the type of object clause according to classification given in Silić and Pranjković (2005). Finally, the whole construction (clause or coordination of clauses) behaving as an object will be enclosed in <OBJ> tag.

4 Object clauses

Function of object clauses in Croatian language is the same as in probably all Indo-European languages; they refer to their superordinate-clause predicate as a direct object. This makes them syntactically dependent on the main clause since they can not behave as stand alone sentences on their own. All types of object clauses have to be preceded by a transitive verb in an active voice form.

The interesting thing about dependent clauses in Croatian language is that it is not possible to predict their function in a sentence by observing only their structure. Function of a clause can be determined only by analyzing its co-text and context. In the following example we will show how the same clause can have different functions in the sentence depending on the context:

- *Vidio sam [da se igra]. – **object** clause*

3.1 (I saw [that he is playing].)

- *Vidio sam ga [da se igra]. – **adjective** clause*

4.1 (I saw him [playing].)

- *Izišao je van [da se igra]. – **purpose** clause*

5.1 (He went out [to play].)

Object clauses can also be easily confused with subject clauses which refer either to the nominal predicate or verbal predicate in passive voice forms.

- *Poznato je [da pušenje uzrokuje rak].*

6.1 (It is well known [that smoking causes cancer].)

- *Kaže se [da je bolje spriječiti nego liječiti].*

7.1 (It is said [that it's better to be safe than sorry].)

According to Silić and Pranjković (2005), three subtypes of object clauses can be differentiated: relative, interrogative and declarative object clauses.

4.1 Relative object clauses

Relative object clauses are introduced by relative pronouns and adjectives as complementizers.

- *Jeste li našli [što ste tražili]?*

8.1 (Have you found [what you've been looking for]?)

- *Kupit ću [kakvog nađem].*

9.1 (*I will buy [the kind I find].)

4.2 Interrogative object clauses

Interrogative object clauses can be divided in seven groups according to their meaning:

1. **general** are introduced by interrogative conjunctions *'li'* and *'da li'* or by interrogative pronouns (*'tko'*, *'koji', 'čiji', 'što', …*)

    - *Još ne shvaćaš [što se dogodilo].*

    10.1(You still don't understand [what happened].)

    - *Zaboravio sam [koji je danas dan].*

    11.1(I forgot [which day it is].)

2. **of place** are introduced by interrogative adverbs of place.

    1. *Recite [kamo ste se zaputili].*

    12.1(Tell us [where you are headed].)

3. **of time** are introduced by interrogative adverbs of time.

   - *Nisu rekli [kad će doći].*

   13.1(They didn't say [when they'll be coming].)

4. **of manner** are introduced by interrogative adverb *'kako'*.

   - *Još nismo saznali [kako se to dogodilo].*

   14.1(We still haven't found out [how that happened].)

5. **qualitative** are introduced by interrogative adjectives *'kakav', 'kakva', 'kakvo'*.

   - *Ne znam [kakav si ti to čovjek].*

   15.1(I don't know [what kind of a person you are].)

6. **of amount** are introduced by interrogative adverb *'koliko'*.

   - *Znaš li [koliko si već popio]?*

   16.1(Do you know [how much you drank already]?)

7. **of cause** are introduced by interrogative adverbs of cause or prepositional expressions *'zašto'*, *'zbog čega'*,

   - *Ne razumijem [zašto si zakasnio].*

   17.1(I don't understand [why you are late].)

4.3 Declarative object clauses

Declarative object clauses are introduced by conjunctions *'da'*, *'kako'* and *'gdje'*, among which *'da'* is the most common. *'Kako'* is somewhat less frequent and appears as a stylistic variant of *'da'*. *'Gdje'* is extremely rare and its use is very stylistically marked.

- *Obećao si [da ćeš doći].*

18.1(You promised that you'll come.)

- *Rekli su [kako ga nije briga].*

19.1(They said that he doesn't care.)

5 Grammar

Grammar that we have composed for this purpose can be divided into four parts. The first part (Figure 1) describes the predicate. We search for a verb phrase in an active voice form. To ensure that the predicate requires the object complement, we are using the information about the verb valency from the lexicon (Vučković *et al.* 2010c). In that way, only those verb phrases that require the complement in accusative case <VP+DCobl=Acc|0Acc|0DAcc|DAcc> will be taken into consideration.

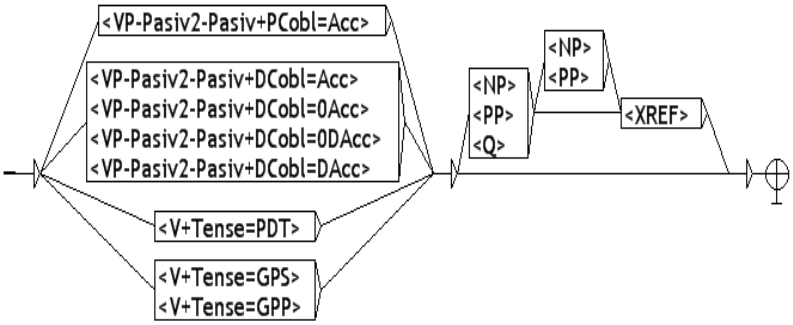


Figure 1. Recognition of the predicate

Croatian complex verbal forms can be split by a prepositional <PP> or noun phrase <NP> or some particles <Q>. This has also been anticipated with the grammar (Figure 1).

Apart from these cases where object clause is the complement of the predicate, we shall include the possibility that it can be a part of a *'predikatni proširak'* (predicate extension), adverbial phrase which describes the circumstances under which the action denoted by a predicate was performed. In these cases, the object clause is preceded by a verbal adverb <GPS> or <GPP>, or passive participle <PDT>.

The subgraph shown in Figure 2 describes everything that can come between the predicate and the related object clause; indirect object, prepositional or adverbial phrases, reflexive pronouns, particles.

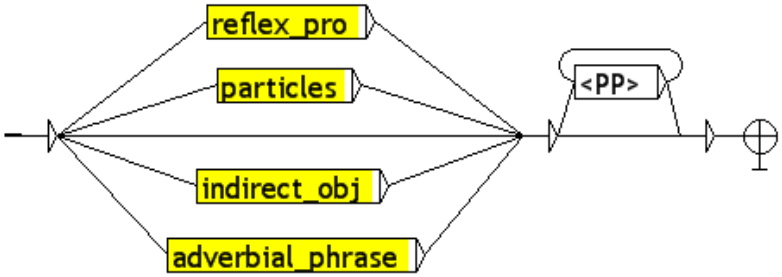


Figure 2. Between the predicate and the clause

In the case of an interrogative object clause, the relating object clause usually immediately follows the predicate. In that case, this part of the grammar is skipped.

Description of object clauses begins in the third part (Figure 3), starting from the conjunctions and complementizers which can introduce the clause. As for describing the clause itself (Figure 4), we did not use any syntactical structures, but kept the definition at the level of words and punctuations. We have already said that the structure of the clause does not tell us much about its function in the sentence.

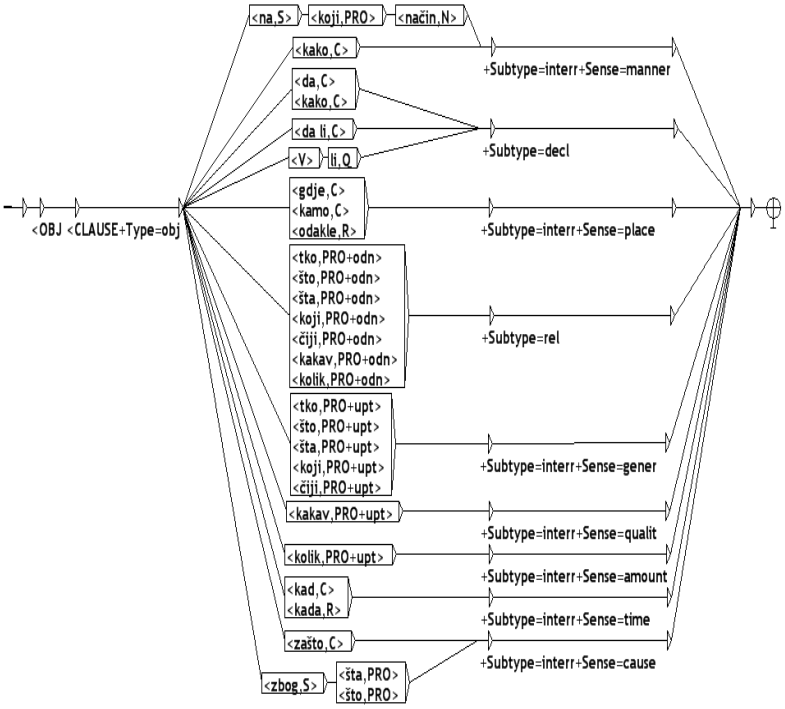


Figure 3. Conjunctions and complementizers introducing the object clause

Only the sequence recognized by this part of the grammar will be annotated.

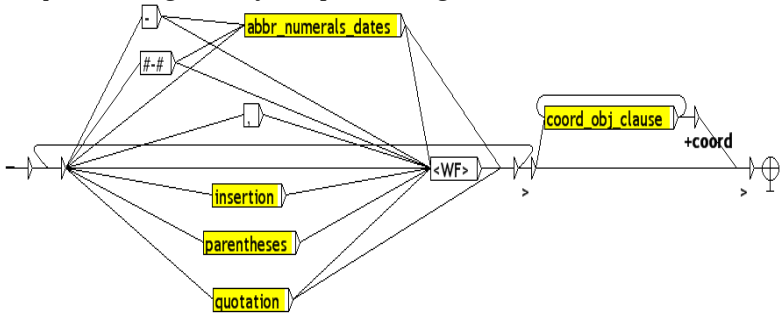


Figure 4. Sentence parts inside the object clause

The fourth part (Figure 5) describes sequences that can occur after the object clauses.

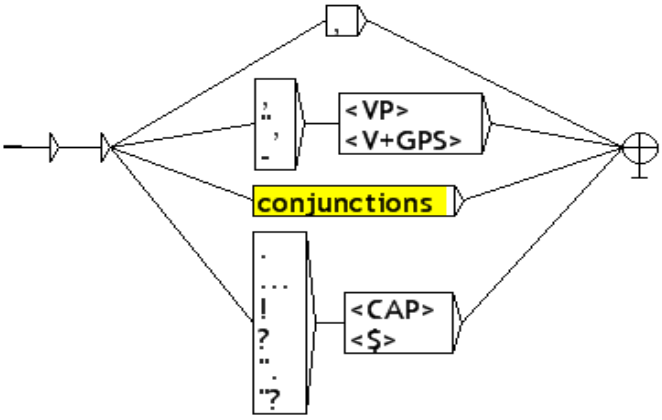


Figure 5. Sentence parts that may appear after the object clause

6 Examples

We shall give few examples of complex sentences in which object clauses can be identified using our grammar. The predicate from the main clause is double underlined for its easier recognition.

[S1] *Dodao je ([da približavanje Hrvatske EU ima dvije faze].*

In the first example sentence [S1] the object clause comes after the predicate. This is the simplest and most common case.

[S2] *Pretpostavimo ([da imate visoke demokratske standarde], [da manjine imaju puna prava], [da su medijske slobode savršene])...*

In [S2], just like in [S1], object construction comes after the predicate. The only difference is that in this example we are dealing with a coordination of object clauses.

[S3] *Zato savjetuje svima koji namjeravaju podići kredite ([da malo pričekaju, ako to mogu]).*

The third example [S3] is little more complex. Between the predicate and the object clause we have indirect object (*"svima"*) and an adjective clause reffering to it (*"koji namjeravaju podići kredite"*).

[S4] *Odgovarajući na pitanje hoće li na dogovore iz Mokrica djelovati skorašnji slovenski lokalni izbori, Maštruko je rekao ([kako u to ne vjeruje] te [da bi u slučaju kad bi države svaki put čekale ([da prođu izbori]), pregovaranje bilo nemoguće].*

[S4] shows the most complex example with several levels of subordination. On the first level we have a coordination of object clauses coming after the main clause predicate. The second clause in the coordination is a complex clause consisting of a main clause (*"da bi u*

slučaju […] pregovaranje bilo nemoguće") and an adjective clause (*"kad bi države svaki put čekale da prođu izbori"*) referring to the prepositional phrase (*"u slučaju"*). That adjective clause is also complex and consisting of a main clause (*"kad bi države svaki put čekale"*) and an object clause (*"da prođu izbori"*). The predicate which that clause is referring to (*"bi […] čekale"*) is split by the subject (*"države"*) and adverbial phrase (*"svaki put"*).

7 Problems

The main problem we were faced with is associated with the detection of the predicate. As Croatian is extremely flectional language, most of syntactical relations are encoded in morphology that results in almost free word order. Because of that, split predicates, or better to say, verb phrases are sometime hard to identify. For instance, in the case of complex tenses, copula can occur in different places in the sentence apart from its main verb. Also, in the case of verb phrases with modal verbs, the main verb is often apart from the modal verb and the copula. Although this has already been discussed by Vučković *et al.*(2010b) and lot of improvement has been achieved, we still don't have 100% accuracy in identifying these cases.

The second problem is related to ambiguity of verb phrases. Certain verb phrases, most often those containing the reflexive pronoun, are often recognized as both in passive and active voice. And we have already mentioned that the predicate has to be in active voice forms, otherwise it cannot have object clause as an argument. This ambiguity is the reason for most of the false-positive matches.

The third problem concerns the verbs' valency frame. We have already explained that, in order to be able to take an object clause as an argument, the predicate must consist of a verb which can take the argument in accusative case. But what if the verb can take two arguments in accusative case? In our valency frame these cases are not recognized. Such are, for example, verbs *'pitati [koga] [što]'* (to ask [whom] [what]) and *'učiti [koga] [što]'* (to teach [whom] [what]). Croatian grammars are not consistent in defining this case: Težak and Babić (2005) don't give any explanation, Barić *et al.*(2005) think that these two arguments are both direct objects while Silić and Pranjković (2005) make the distinction between them as one being direct and the other indirect object, but also give not very convincing explanation as to which is which.

However, the argument in accusative that can be replaced with an object clause will always come second. This means that we have a situation in which a potential object clause is preceded by another object, i.e. noun phrase, in accusative case. That kind of construction is typical for adjective clauses and, therefore, can not be allowed since we would have very large number of false-positive matches. To solve this problem, we shall have to provide additional description for such verbs in the valency frame and create a special path in the first (Figure 1) and second part (Figure 2) of the grammar.

The last two problems are related to phenomena beyond syntax and therefore impossible to deal with at this level of language analysis. One of them is the problem of identifying the level of subordination of a clause which is a problem of semantics. And the second problem concerns the orthographical rules of Croatian, especially of using punctuation markings like comma and dash since comma is sometimes used for prosodic and not only logical reasons while dash can be used as colon, semicolon, quotation mark, etc.

8 Evaluation

Bearing in mind all aforementioned problems, we decided to perform the evaluation in ideal circumstances in order to obtain the scores dependent only on this model, ignoring all other factors. Ideal circumstances imply that in all of our test examples, verbal phrases serving as predicates are correctly identified (i.e. chunked) and that the information about the verb valency is present.

Our test corpus consists of 174 complex sentences with 215 object clauses. The model scored an overall F1-measure of 0.59 as shown in Table 1.

Precision	Recall	F1-measure
0.46	0.82	0.59

Table 1. Evaluation scores

Since the grammar is designed to work in *all matches* mode, low precision was expected. However, the correct result was present within the returned matches in 91% of the cases. Depending on the complexity of the sentence, the number of returned matches ranged between 1 and 12 and the average number of returned matches per clause was 2.15. It is evident that some kind of disambiguation will have to be performed on the matches but that will be possible only when all (or at least the most frequent) types of clauses in Croatian will be described in this way (Vučković *et al.*, 2010a).

We believe that relatively high recall, on the other hand, confirms the adequacy of the model. We can also expect somewhat better results in the future since we have identified the critical cases that will enable us to work on the improvements of the model.

9 Conclusion and future work

The paper describes a model for the recognition and annotation subcategory of dependent noun clauses known as object clauses. This first try in Croatian clause segmentation gives promising results and as such serves as an introduction into clause detection and sentence classification for Croatian texts in general (Vučković *et al.*, 2010a).

Our future work will include solving the problems described in Section 7 that will lead to higher precision and recall of the model. Improved model may bring us closer to deep parsing of Croatian as well.

Keynotes: clause detection, Croatian language, NooJ, object clauses, partial parsing.

Acknowledgments
This work was done within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grant 130-1300646-1776.

References

Barić, E. et al., 2005. *Hrvatska gramatika* 4th ed., Zagreb: Školska knjiga.

Boras Damir. 1998. *Teorija i pravila segmentacije teksta na hrvatskom jeziku*. PhD Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Ejerhed, E.I., 1988. Finding clauses in unrestricted text by finitary and stochastic methods. In *Proceedings of the second conference on Applied natural language processing*. Austin, Texas: Association for Computational Linguistics, 219-227.

Leffa, V.J., 1998. Clause processing in complex sentences. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain, 937–943.

Orasan, C., 2000. A hybrid method for clause splitting in unrestricted English texts. In *Proceedings of ACIDCA '2000, Corpora and Natural Language Processing*. Monastir, Tunisia, 129 - 134.

Ram, R.V.S. & Devi, S.L., 2008. Clause boundary identification using conditional random fields. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*. Haifa, Israel: Springer-Verlag, 140-150.

Silberztein, M., 2003. NooJ manual. Available at the web site http://www.nooj4nlp.net (200 pages).

Silić, J. & Pranjković, I., 2005. *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*, Zagreb: Školska knjiga.

Težak, S. & Babić, S., 2005. *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje* 15th ed., Zagreb: Školska knjiga.

Vučković K., Agić Ž., Tadić M. 2010a. "Sentence Classification and Clause Detection for Croatian". In M. Tadić, M. Dimitrova-Vulchanova, S. Koeva (eds) *Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages*, Croatian Language Technologies Society, Faculty of Humanities and Social Sciences, Zagreb, 131-138.

Vučković, K., Bekavac, B. & Dovedan, Z., 2010b. SynCro - Parsing Simple Croatian Sentences. In A. Ben Hamadou, S. Mesfar, & M. Silberztein, eds. *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*. NooJ 2009 International Conference and Workshop. Touzeur, Tunisia, 207-217.

Vučković, K., Mikelić Preradović, N. & Dovedan, Z., 2010c. Verb Valency Enhanced Croatian Lexicon. In T. Varadi, J. Kuti, M. Silberztein (eds) *Applications of Finite-State Language Processing - Selected Papers from the2008 International NooJ Conference*. Cambridge Scholars Publishing, Budapest, 52-60.

Vietnamese classifiers processing for nominal syntagms extraction

Hồ Đình Océane

CRIM, INaLCO
Paris, France.

Abstract

This paper presents the implementation of Vietnamese classifiers in the nominal syntagm modelization. This work has been done in the context of Viet4NooJ module development, the NooJ module for Vietnamese language, currently under construction. After a linguistic description of those words and a presentation of the resources that have been developed according to it, we shall analyze the first results and propose solutions to improve the system.

1 Introduction

Vietnamese language is an isolating language and, as the other non inflectional languages, has recourse to function words in many cases. In Vietnamese, a function word class is used in noun determination: the classifiers. They carry useful informations for text analysis tasks. As they are one of the main components of nominal syntagm, processing them is essential and preliminary to textual disambiguation task.
In the first part we will present nominal syntagm in Vietnamese sentence, then the classifiers and their function in noun determination. We will explain what kind of informations they carry are useful in natural language processing.
In the second part we will detail the resources we set up and how they have been implemented in the Viet4NooJ module. We will present the graphs modelling the nominal syntagm and the lexical resources produced for our work.
Then, in the third part, we will see how classifiers identification appeared as problematic in the nominal syntagm extraction task and the questions risen concerning disambiguation classifier / noun and dealing with cases of classifiers presence / absence – either compulsory or allowed.
Finally, in the fourth part, we will expose the choices we made for making the most of the informations carried by the classifiers without loosing quality in the nominal syntagm analysis.
We will finish with a discussion on what perspectives remain to be explored to improve the system.

2 Classifiers in Vietnamese language

Before presenting the classifiers themselves, some indications about Vietnamese language and the Vietnamese nominal syntagm structure can be useful.

2.1 Vietnamese language

Vietnamese is a non inflexional language. Consequently, all the functions that are taken on by morphology in flexional languages will be here taken on by various tool words. These functions are very numerous and varied. For example, differentiating grammatical classes, but also marking tenses, gender, number, etc. We can represent classifiers as one of those classes of tool words, operating in the nominal syntagm. That is why we first need to describe the structure of nominal syntagm (NP) in Vietnamese. NPs are built following the pattern described here :

NUM + CL + N + QLF + DEM

NUM stands for numeral, which means here any expression of number or plural. As we can see, classifier precedes the noun, when qualifiers come after, and at last come the demonstratives.

That said, the first difficulty appears as it seems hard to differentiate classifiers from nouns in two distinct closed classes. Word forms that appear in classifier position can also be found as substantives, often with a more or less different meaning. In order to overcome this difficulty we shall now take a deeper look on what are the classifiers.

2.2 Vietnamese classifiers

A first distinction can be made between the 2 generic classifiers on one hand and what we can call the specifiers on the other hand. The two generic classifiers are used to mark the main distinction made between nouns: the feature of animacy. The word “cái” precedes inanimate nouns while the word “con” precedes animate nouns. Concerning specifiers, they can be used to distinguish the parts of a vegetable:

1a. “một quả chanh”
one SPEC[fruit] lemon “a/one lemon”
1b. “một cây chanh”
one SPEC[tree] lemon “a/one lemon tree”
1c. “một lá chanh”
one SPEC[leaf] lemon “a/one lemon leaf”
1d. “một hạt chanh”
one SPEC[seed] lemon “a/one lemon seed”

But the same words can also denote the shape of an object: the noun “quả” (“fruit”), used as the classifier for fruit, is also used with round objects; “cây” (“tree”), used as the classifier for trees, is also used with elongated objects; “lá” (“leaf”), used as the classifier for leaves, is also used with flat and thin objects; “hạt” (“seed”), used as the classifier for seeds, is also used with small and rounded objects:

2a. “quả đất”
SPEC[round object] earth “Earth”
2b. “một cây bút”
one SPEC[elongated object] pen “a/one pen”
2c. “một lá cờ”
one SPEC[flat&thin object] flag “a/one flag”
2d. “một hạt mưa”
one SPEC[small&rounded object] rain “a/one rain drop”

2e. “một hạt trai”

one SPEC[small&rounded object] oyster "an/one oyster pearl"

In the examples 1(a to d) the function of the classifiers is easy to understand: for a same word form, for example here "chuối", a vegetable species, they will make possible to distinguish the fruit from the tree or the leaf. But in the examples 2(a to e), it is harder to admit the necessity of the presence of this word form preceding the noun, giving an inherent property of this noun. If this property is already inherent to the nouns why would we have to add another word expressing it? Moreover, note that in 2(a to c) it has totally disappeared in the translation, when in Vietnamese its omission won't be allowed.

2.3 Classifiers functions

This is because the classifiers stand for other functions in the NP. As we already saw, they can play semantic function, specifying a part of a vegetable. They also can cumulate both semantic and syntactic functions.

Quantificabilisation and reification: In Vietnamese, nouns are uncountable and they denote abstract notions. Therefore, classifiers are needed to allow nouns to appear with a numeral expression, or to denote definite occurrences, individualized out of the notion denoted by the following noun. They are used to come from an abstract notion to a concrete object that will then be able to be more specified, if needed.

Measure: Classifiers can also be used to express measure. The process is the same, as we want to bring out unities, that will then be able to be counted, qualified or designated, from a set, except that in Vietnamese it has to be done with all the nouns, when in English, it is only necessary for the class of nouns that denote mass notions (water, flour, etc.). Specify the way to measure is needed, with terms like "a glass of", "a liter of", etc. In Vietnamese the distinction between mass nouns and entity denoting nouns is not that relevant, as in both cases specifying which unities are to be enumerated is needed.

3 Resources implemented for NooJ

As we saw earlier, classifiers mobilize both syntactic and semantic levels. That is why the resources implemented in the Viet4NooJ module to analyze classifiers gather both dictionaries and syntactic grammars (represented by graphs), in order to take into account both lists of word forms and syntactic positions.

The classifiers are grouped in a 143 entries dictionary, based on the list made by Nguyễn Phú Phong[1], with the tag SPEC.

These entries are then retrieved by a grammar modelizing the nominal syntagm. This grammar is made up of a main graph and two sub-graphs: one for the classifier-noun set, and another one that has been formerly developed for a Named Entities application and that we included in the nominal syntagm modelization.

[1]Nguyễn, 1995, p.111-113

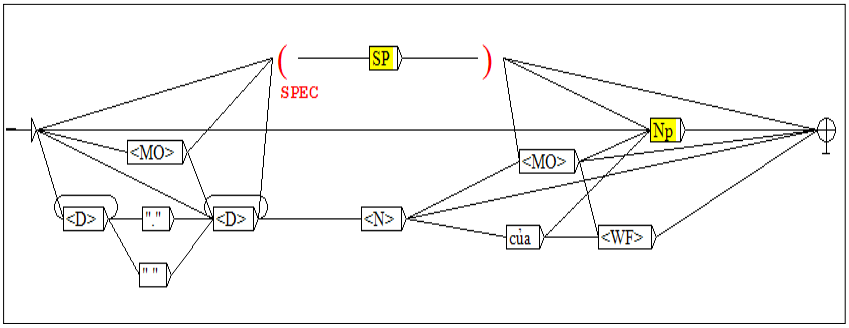


Figure 1. Nominal syntagm, the main graph

The group of states seen on the left, before the first sub-graph, represent complex digits. It can be noticed that this first sub-graph can appear without any expression of number before, as this case can happen.

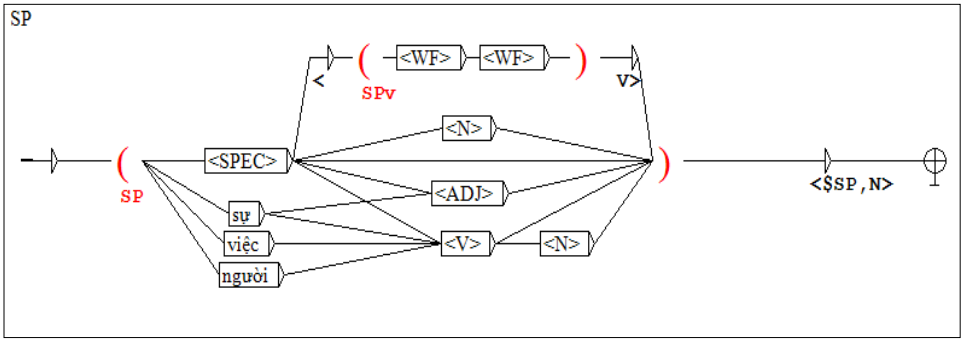


Figure 2. Sub-graph modelizing the classifier-noun set

We can see that classifiers (here SPEC) can appear before other categories than noun: these are nominalization cases.

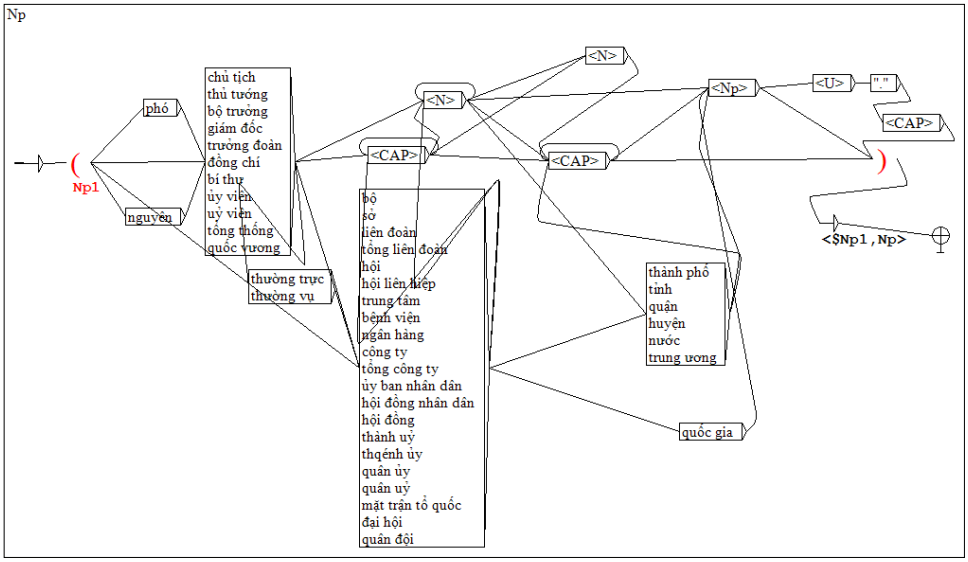


Figure 3. Sub-graph modelizing proper nouns

As it can be seen in the graphs, some other grammatical categories are also used: <N>, <V>, <ADJ>, <WF>, <CAP> or <D> correspond to the generic categories in NooJ, and to the tags used in Viet4NooJ module's dictionaries. But we can notice that a new category has been added: <MO>, for the category of words that has been discussed earlier: tool words.

4 Questions risen

The first results let appear a lot of noise. The main difficulty was to determine the border of the nominal syntagm: where does it starts, where does it finishes. Locating tool words was a good indication in that task, but the <MO> category covers too many and too different cases and should be detailed.

We also miss complex cases of nominal syntagms, as expression of possession or coordination, which deserve a better and more detailed description.

Another problem remains that a same form can be either a classifier or a noun. The border between the two categories is very porous. This is due to the fact that all classifiers are former nouns that have lost a part of their meaning to become classifiers. And this phenomenon is possible with almost any of the nouns. Choices have to be maid for the task of establishing a list of what word form we consider as classifiers.

And to add even more difficulty, many, if not all, syntactic positions can be empty, of course not all at the same time, but all cases of omission can occur according to different situations.

5 Considering solutions

The first thing has been to completely remove the <MO> tag from the nominal syntagm grammar and to split the tool words list into smaller lists with more precise tags. For example, the only category of tool words that can appear before the classifier is plural markers (“các”, “những”), in the same syntactic position as digits, or in combination with

them. Letter versions of numbers have also been added. Similarly, demonstratives ("này", "kia") can only appear at the end of the nominal syntagm. They are even a reliable clue of where the NP finishes.

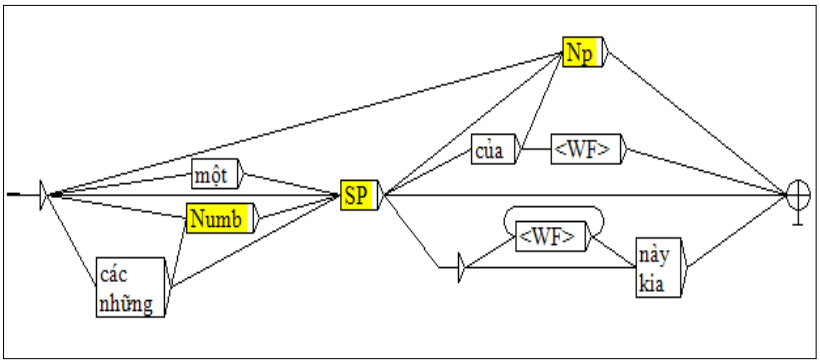


Figure 4 . Illustration of improvements on the main graph

Another function permitted by NooJ is the possibility to handle recursivity. This will be useful to deal with the expression of possession. A possessive nominal syntagm is introduced by the tool word "của". A NP introduced by this tool word comes in determination of another NP, which means inside it: [... [của ...]$_{NP}$]$_{NP}$. This will be possible to deal with, using recursivity function.

Some other external elements, before or after the NP, can also help determining the borders of it. For example verbs can play this role, establishing the beginning of a verbal syntagm and therefore the end of the nominal syntagm. This means enlarging the grammar the the entire sentence.

We also need to study in what cases what syntactic position can be empty, in order to reduce noise in the result, being able to understand what cause permits the omission of an element.

6 Perspectives

Until now, we have mainly focused on the syntactic behavior of the classifier, but we did not yet exploit all the semantic informations given by it. Going always deeper in the description of classifiers, we need to enrich the <SPEC> tag with a set of features, as classifiers play various roles in noun determination. As we saw, they can denote measure, individualization, nominalization, etc. and we need to describe rules of combination between classifiers and nouns, in order to collect more informations during the NP retrieving task.

Acknowledgments

I would like to thank Max Silberztein for his technical help, reactivity and enthusiasm, and Philippe Lambert for introducing me to NooJ, permitting me to integrate the Viet4NooJ development and, above all, for his precious support at all times.

References

Nguyễn Phú Phong. 1995. “Questions de linguistique vietnamienne. Les classificateurs et les déictiques”. Presses de l’Ecole française d’extrême-orient, Paris, France.

Löbel, E. 2001. “Classifiers and semi-lexicality: Functional and semantic selection”.In N. Corver & H. van Riemsdijk (eds.), *Semi-lexical Categories: The Function of Content Words and the Content of Function Words*, Mouton de Gruyter, Berlin-New York, 223-271.

Emeneau, M. B. 1951. “Studies in Vietnamese (Annamese) Grammar”. *University of California publications in linguistics*, volume 8, Berkeley, Los Angeles, USA.

Nguyễn Tương Hùng. 2004. “*The structure of the Vietnamese noun phrase*”. Boston University, Boston, USA

Do-Hurinville, D. T. 2008. “Nominalisation et construction du thème en vietnamien”. *Faits de Langues* 30, 209-216.

Truong, Văn Chình. 1970. “*Structure de la langue vietnamienne*”. Impr. nationale, P. Geuthner, Paris

Culioli, A. 1999. “*Pour une linguistique de l'énonciation, Domaine notionnel*”, volume 3, Ophrys, Gap.

Bisang, W. 1999. “Classifiers in East and Southeast Asian Languages : counting and beyond”. In *Changes in Numeral Systems*, Mouton de Gruyter, Berlin, 113-185.

Simpson, A. 2005. “Analyticity in the nominal domain: classifiers and the structure of DPs in languages of Southeast Asia”. LSA Syntactic Analyticity, MIT July

François, A. 1999. “L'illusion des classificateurs”. In *Faits de langues* 14, 165-175.

Vogel, S. 2002. “Détermination nominale, quantification et classification en khmer contemporain”. In *Bulletin de l'Ecole française d'Extrême-Orient*. volume 89, 183-201.

Selection criteria for method of translation and some suggestions for the platform NooJ

Hajer Sahnoun[(1)]and Kais Haddar[(2)]

[(1)]*LaboratoryMIRACL, FSEGS, Sfax 3018, Tunisia*
[(2)]*Laboratory MIRACL, FSS, B.P. 1171, Sfax 3000, Tunisia*

Abstract

The aim of our work is to help NooJ users and designers in the automatic translation (AT) field. This help is in term of assimilation and judging of multiple situations that can be faced by AT researchers. In this paper, we are focusing on presenting the linguistic approach of AT. Then, we perform a comparative study between methods and we define some criteria facilitating the choice of the suitable method. We note that our study isperformed without binding in the performance of existing automatic translators obeying linguistic methods. But, a phase of experimentation using the linguistic platform NooJ is elaborated to allow us the proposition of a number of suggestions to ameliorate the capabilities of NooJ in the AT field.

1 Introduction

The necessity of AT is always in continuing increase. As consequence, many works are achieved touching different subjects relating to AT domain. Researchers are always trying to invent, improve and provide additions to this domain. In fact, the importance of this domain and its difficulties makes searching about it a relevancy.

To create an automatic translator, we need linguistic and technical knowledge. We focus our study in technical knowledge that acquires good study about approaches and methods of AT. The complexity of AT domain makes the elaboration of studies a necessity to put beginners in the heart of the domain. In general, it's not easy for a beginner to acquire knowledge about AT because of multiplicity of documents and terminologies.Also, the mission of deciding the suitable method concerning the objective under the creation of automatic translator is complicated. Deciding the use of a method is not arbitrary that's why it's a relevance to find a mannerto judge the necessity of a method to a well defined situation. To take the decision, it's necessary to compare between methods and to find a way to decide the more contributable one. Furthermore, the choice of the manner of building an automatic translator can repose on a programming language or a linguistic platform. In fact, it's important to study the capabilities of a linguistic platform before its use.

In the literature, we note the inexistence of works comparing the effect of use of every method. All works of comparative study compare results of some automatic translator like in Babych (2007) in order to compare the effect of use of every method without taking on consideration some criteria like the goal of build etc.

The objective of our work is to elaborate a comparative study between linguistic methods that allows us the detection of some criteria facilitating the choice of the suitable method.

Also, we use the linguistic platform NooJ in order to judge the use of a platform on the build of an automatic translator.
In this paper, we begin by an overview representing the AT domain. Then, we describe our comparative study and we propose a set of criteria that we define to facilitate the choice of the suitable method. After that, we experiment the applicability of linguistic methods of AT using the linguistic platform NooJ (Silberztein: 359-370) in order to judge its position in the AT domain.

2 Overview

From the beginnings of first efforts in the AT field and until today, many works concerning AT are appearing. These works permit the building of a number of AT systems like in Claveau (2007) and Wehrli (2007). According to (Sahnoun: 257-270), by focusing on works done with the linguistic platform NooJ that we study, we cite the work presented in Papadopoulou and Gavriilidou (2009) and in Fehri (2009). In fact, these examples present just a number of AT systems which demonstrates the importance of efforts in the domain of AT.
Concerning the linguistic approach that we interest, it considers different linguistic phenomena in order to achieve satisfactory results.Indeed, it describes a set of methods constituting a rich linguistic amalgam.In fact, the different paths of the famousVauquois's triangle illustrated methods belong on this approach and the depth of analysis of each. According to (Dorr: 1-20), the following figure illustrates the various methods and describes their linguistic levels of analysis:

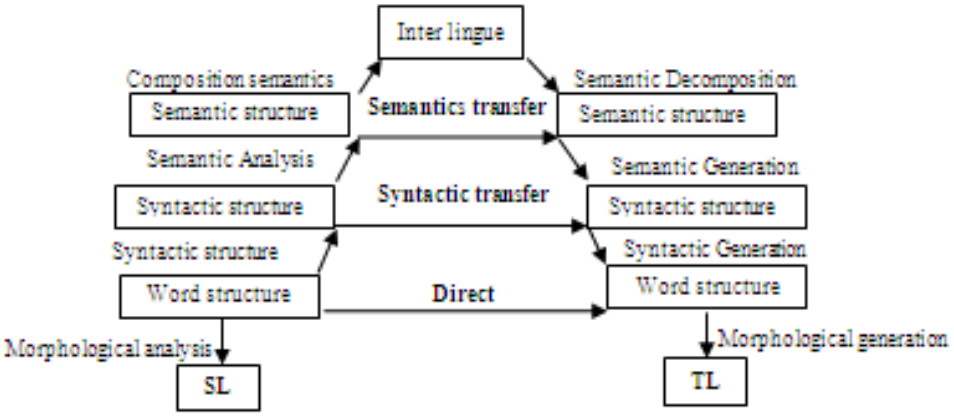


Figure 1. Triangle of V auquois for AT

Figure 1shows that the linguistic approach can be divided into three major methods each one has a number of characteristics. For us, we add another method that we consider trivial. This method is the semi-direct method. From here, we consider the direct, the semi-direct, the transfer and the pivot based method.

3 Comparative study between linguistic methods

The comparative study between linguistic methods of AT consist todefinition of a set of points contributing to distinct between capabilities of each method. These points affect the

right position to every method. So, we define the use of resources, techniques used, level of analysis, evaluation of quality as criteria of comparison.

3.1 Use of resources

The main resources used by the linguistic translation are essentially dictionaries and grammars. The dictionaries present the lexical database on what the process of AT is based. Thus, the grammars use multiple rules to manipulate the lexicon.We consider as another resource that we called heuristics. Heuristics describe transformation and filtering rules. Also, heuristics are not formalized as grammars and have superficial definition. The following table describes the use of resources by each method:

Methods / Resources	Direct	Semi direct	Transfer	Pivot
- Dictionaries				
- Monolingual		**X**	**X**	**X**
- Bilingual	**X**	**X**	**X**	**X**
- Multi-target	**X**	**X**	**X**	**X**
Grammars			**X**	**X**
Heuristics		**X**	**X**	**X**

Table 1. Use of resources by linguistic methods of AT

Table 1 shows the indispensability of dictionaries for every method of AT. The direct method use only dictionaries to translate. So that, the degree of resources utility, used by other methods, is distributed between dictionaries, grammars and heuristics.

3.2 Used techniques

A number of techniques is used to ensure the application of linguistic methods of AT. Among these techniques, we cite syntactic projection, attachment, identification of concepts, etc. We summarize in the following table most important techniques and the use of these techniques by each linguistic method:

Methods / Techniques	Direct	Semi direct	Transfer	Pivot
Lexical analysis	X	X	X	X
Syntactic projection			X	X
Attachement			X	X
Use / Construction of pivot language (PL)				X
Hybrid interface structure			X	
Lexical transfer / correspondance	X	X	X	X
Structural transfer			X	X
Readjustment		X		
Concepts and relationships			X	X

Table 2. Relation betweenmethods and techniques of linguistic AT

Table 2 shows that the lexical transfer is essential for all methods. The direct method uses simply lexical analysis and lexical transfer to ensure the translation. Concerning the semi-direct method, it uses readjustment as technique that attributes to this method more precision comparing with the direct method. The transfer method uses wholly or partially a number of techniques such as lexical analysis and attachment. The pivot method can use all techniques especially the transfer method for analysis and generation. Systems using the pivot method can use a LP existing or already defined.

3.3 Level of analysis

The reached analysis level differs depending on the used translation method. All methods get a definite depth of analysis.According to this depth, the quality produced can vary especially if structures to translate are complicated.Indeed, we present the table below, to illustrate the analysis level reached by linguistic methods:

Methods / Level	Direct	Semi direct	Transfer	Pivot
Morphologic	X	X	X	X
Morphosyntaxic		X	X	X
Syntactic			X	X
Semantic			X	X

Table 3. Levels of analysis affected by linguistic methods of AT

Table 3 shows that all linguistic methods affect the morphological level without exception. In fact, the direct method uses only this level which reflects the superficiality of analysis achieved by this method. Moreover, the semi direct method pushes analysis to morpho-syntactic level. The transfer and the pivot based method are able to go further in terms of depth of analysis. In fact, analysis used by the transfer method varies. Indeed, it can affect the syntactic level only or combined it with semantic analysis. Obviously, even syntactic analysis employed may have different levels such as constitutional and functional analysis. While the pivot based method reached immediately the semantic level.This exceeds the

simple semantic analysis to the definitionof semantic representation or an intermediate language.

3.4 Quality evaluation

The evaluation of quality of each linguistic AT method allows us to deepen more our comparative study. This evaluation is mostly about calculating the similarity between translations of human experts and translations produced by AT systems. In fact, a set of measures such as Blue, NIST, WMR is needed also some criteria such as fluency and adequacy can be used. While looking for a way to distinguish between linguistic methods in terms of evaluation, we believe in finding methods and / or measures of evaluation that concern some methods and not other. We note that there are no measures or formula of evaluation made especially for the evaluation of linguistic methods. In order to compare qualities of translations produced by linguistic methods, we attribute a class of quality to each one as illustrate the following figure:

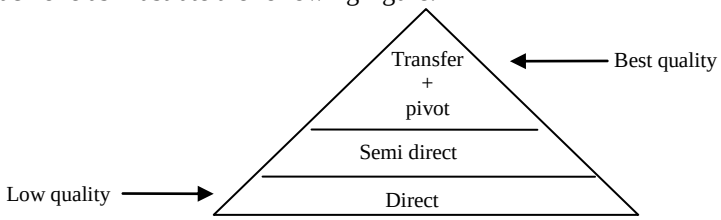


Figure 2. Level of quality of linguistic methods of AT

Figure 2 assigns to each linguistic method a position reflecting the quality of translation produced. The summit of the triangleis attributed to the best quality of translation while its baseindicates the low quality. Therefore, we are positioning transfer method and the pivot based method at the summit of the triangle. These methods have the capacity to produce higher quality of translation. However, the direct method takes the position in the base of the triangle because it provides the lowest quality. The semi direct method has an intermediate position due to the average quality that produce.
We continue our task of clarification by directing our interest to the choice of the suitable method depending on the situation. So we focus on defining a process of choosing the appropriate method according to a set of criteria that we judge essential.

4 Proposition of selection criteria

In order to design and implement an automatic translator, it's necessary to take into account a number of criteria facilitating the choice of the most suitable method. This avoids providing extra effort useless for translation or providing superficiality effort for translator demanding in terms of depth of analysis.
In the following, we define and discuss a number of criteria that influence the choice of the linguistic method. Then, we present each criterion and we describe his relationship with different methods.

4.1 Aimed domain

The aimed domain of the automatic translator under construction influences greatly the choice of the linguistic method. In fact, we consider the domain taking accountof both sides. The first side describes the generality of the domain that implies the possibility of covering a specific domain or a number of domains. The second side describes ambiguity of the lexicon describing the targeted aimeddomain. Indeed, the ambiguity generated by the lexicon increases the degree of difficulty in interpreting the meaning of structures to translate. The following table illustrates the influence of the aimed domain to the choice of suitable method of AT:

Methods / Domain	Direct	Semi direct	Transfer	Pivot
Specific domain + unambiguous lexicon	X	X		
Specific domain + ambiguous lexicon			X	X
Multi domains			X	X

Table 4. Choice of the method according the criterion « aimed domain »

Table 4 illustrates the choice of the suitable method of AT according to variations of the criterion "aimed domain". A specific domain describes a well determined sub language. When it's about this type of domain, the choice of the method of AT varies with the ambiguity of the lexicon of the domain. However, a specific domain combined with an unambiguous lexicon does not need deep analysis. In consequence, the direct method and the semi-direct method, non-demanding in terms of analysis, are sufficient. However, a specific domain containing an ambiguous lexicon increases the number of linguistic problems. Indeed, this type of domain requires the use of the transfer or the pivot based method. If it's about a set of multiple domains, the number of linguistic conflicts increases. Thus, the multiplicity of domains take in account by the automatic translator can create multiple interpretations of meaning for the same terms and structures. So, the transfer and the pivot based method are most appropriate methods to the multiplicity of domains.

4.2 Type of structure

The type and the complexity of the structure to be translated can guide the use of a well-determined method. An automatic translator can target the translation of simple lexical units or structures with a varying degree of complexity. The following table illustrates the distribution of translation methods by the type of structure to translate:

Methods / Structure	Direct	Semi direct	Transfer	Pivot
Word	X	X		
Phrase		X	X	X
Sentence			X	X

Table 5. Choice of AT method according to criterion « type of structure »

Table 5 shows that when the complexity of structure to be translated, relations between terms and relationshipsbetween structures increase, the requirement in terms of analysis increases. Indeed, we consider that the direct and the semi-direct method are most suitable for the translation of simple words, because of lack of relations with other terms. In this case, the translation of words can be solved by the simple use of bilingual dictionaries like the translation of names of vitamins. On the other hand, in case of necessity of certain analysis that affects to words the proper form of translation in TL, the semi direct method is to choose like in Fehri (2009) presenting the translation of names of sports places.Concerning phrases, the semi direct method is also sufficient given the simplicity of structures and the possibility to identifythe multitude of construction but it's possible to use the transfer and the pivot based method. These two methods refine more the result of translation given analysis that they provide. Concerning sentence structure, they are complex and can have a different construction and grammar form, that's why the direct and the semi direct method can never serve to their translations. Indeed, the transfer and the pivot based method are most suitable for translation of this type of structure. More the structure to be translated by the system is complicated more the use of a linguistic method that has better capabilities in term of analysis and treatment is a necessity.

4.3 Type of system

Among criteria that must be taken into account when designing an automatic translator, we cite the number of languages taken into account by the system. In fact, the system can cover a couple or a set of languages. Thus,there are bilingual and multilingual automatic translators. The following table illustrates the use of methods according to the type of system:

Methods / System	Direct	Semi direct	Transfer	Pivot
Bilingual	X	X	X	
Multilingual	(X)	(X)	X	X

Table 6. Choice of ATmethod according to criterion « type of system »

Table 6 shows the influence of the type of system to build on the choice of the linguistic method. Indeed, a bilingual translator takes in consideration a couple of well-defined languages and performs the translation in a specific way. Conceptually, the direct and the semi direct method don't provide any representation resulting in the analysis phase.These two methods of translation are performed sequentially in order to procure a result without providing any module that can be reused. Thus the translation process performed by the direct and semi direct method takes in consideration the direction of the translation that's why these methods are suitable in most cases for bilingual systems.Concerning the transfer method, it's suitable for bilingual translation and it provides a level of analysis that can reach the semantic level. Moreover, the difficulties of the pivot based method are tolerated to make multilingual translators. Given the complexities associated to the pivot method, the transfer method is considered as the most suitable especially because it can be used to make bilingual and multilingual systems. Also, we note that there are limited numbers of cases in which the direct and the semi direct method are useful for multilingualism. These cases are illustrated by the general principles that we derive in what follow.

4.4 Need of use

The need of use of an automatic translator varies with users' linguistic knowledge. This comes from the variation of assimilation levels of users. In fact, more than the level of users' assimilation increases, more they can understand the result produced by automatic translators even if it's not sufficiently. The table above illustrates the principal needs and combines them with adequate methods:

Methods / Need	Direct	Semi direct	Transfer	Pivot
Degree of relevance	X	X		
Identification of context	X	X		
Obtaining surface structures			X	X

Table 7. Choice of ATmethod according to criterion « Need of use »

Table 7 describes the influence of the need of use on the choice of the linguistic AT method. The user can use a translator to decide the relevance of document that he need. Thus,browsing a translated document in order to judge its relevance does not require a good quality of translation that's why the direct and the semi-direct method are sufficient.In addition, the need of the user can be the identification of the context. This identification is in favor of several types of applications such as information retrieval and arrangement of documents. The direct and the semi-direct method are also sufficient for the identification of context. However, the need in term of understanding requires obtaining of surface structures. In fact, the deduction of these structures is performed by deep analysis. Thus, suitable linguistic methods are the transfer and the pivot based method. If the level of understanding that needs the user is high, the translation method chosen must have a capacity to produce results more skilled. Otherwise, even methods which do superficial analysis are sufficient.

4.5 Relationship between criteria

Testing the influence of each criterion independently on the choice of the method of translation is not enough. Indeed, it's essential to take in consideration all criteria to avoid contradictions. To clarify the situation, we begin by presenting a set of definitions and notations. These allow us to identify a set of principles that facilitates the choice of the linguistic method.

4.5.1 Definitions and notations

In trying to resolve the problem of choosing the suitable method of AT, we rely on the definition of four sets. These sets that describe the values taken by each criterion already definedpreviouslyare presented as follows:

Aimed domain= {specific + unambiguous lexicon, specific+ ambiguous lexicon, multi-domain}.

Need of use = {degree of relevance, identification of context, obtaining of surface structures}.
Type of system = {bilingual, multilingual}.
Structure type = {word, phrase, sentence}.

Sets listed above make the definition of principles that describe the choice of the suitable method of AT clearer. Therefore, we present identified principles relying on the collection of values based on four sets defined previously.

4.5.2 Defined principles

The decision concerning the most suitable method depends on the tuple (d, n, t, s) ∈Domain ×Need ×Type × Structure. Thus, we defineeleven principles facilitating the task of makers of automatic translators that we cite a number among them:

P1: The direct method ensures the multilingualism only in translating words.
P2: The semi-direct method ensures the multilingualism only in translating phrases or words belonging to an unambiguous specific domain.
P3: (unambiguous specific domain,n, bilingual, phrase) with n ∈ (degree of relevance, identification of context) →Semi direct.
P4: (unambiguous specific domain, obtaining surface structures, bilingual, sentence) →transfer.

Previous principles present recommendations concluded from our studies and observations.They are neither rules nor obligations. In order to deep more our study, we try to apply linguistic methods using the linguistic platform NooJ.

5 Experimentation and evaluation

A phase ofexperimentation is performed using the linguistic platform NooJ. It was an opportunity to judge the use effect of a linguistic platform to perform translation. Also, it's help us to judge the capacity of NooJ in the AT field. Our observations showed that themajority of works using NooJ like in Barreiro (2008) and in Fehri (2009) uses the semi-direct method.This doesnot encourage constructers of AT systems to use it. That's why we try to implement more complex methods with NooJ.

Our experimentation allows us to offer some suggestions for NooJ contributing to ameliorate its capabilities. We begin by proposing a set of conceptual suggestions and we continue with technical suggestions.

5.1 Conceptual suggestions

Conceptual suggestions care about improving the process of translation. It not concerns the manipulation of NooJ but they describe two main ways that contribute to improve the translation process currently used by NooJ. These two ways are the deepening of analysis and the modulation of the translation process.Analyses performed during the translation done by NooJ were sufficient given the simplicity of structures translated by referring to the results already achieved. However, we must push more analysis in order to deep more syntactic analysis and integrate semantic analysis in the translation process.

Despite the performance of NooJ in terms of syntactic and semantic analysis, the challenge was always to play on the level of description in dictionaries. To apply other methods than

the semi direct, it's important to acquire information belonging to different linguistic levels. This requires gathering more information from additional analysis. This information can help determine components, functions andpredicates informationetc. Thus, we must go beyond the resolution of linguistic problems referred through local grammars, as is currently done by NooJ in favor of deepening analysis. (Silberztein, 177-189) has demonstrated the ability to push NooJ parsing. The following transducers present an example of functional structure detection:

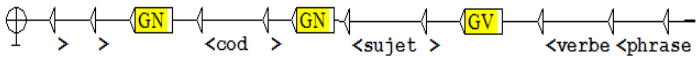


Figure 3. Transducer illustrating a functional analysis

Figure 3 demonstrates the possibility of detection of functional structure using NooJ. Also, NooJ can perform important analysis like syntactic and semantic analysis. Thus, the ability of this platform assigns it a favorableposition in the field of AT.
Concerning the modulation, it consists to decompose the current process of translation performed by NooJ into modules. If the border between modules becomes clear, it will be possible to make multilingual automatic translators. Also, it will be possible to use ready modules, realized by users of NooJ, in several works of translation. In fact, multilingualismcannot be done by the translation method currently provided by NooJ.Indeed, assembly of the translation steps on the same graph deprives researchers to exploit the results of analysis performed in other works.This is due to direct mapping of lexical and analytics information into the desired lexical space without any separation.Also, the modulation facilitates the exploitation of analysis modules already realized that favors the collaboration between the communities NooJ. On the other hand, absence of independent rich analysis modules eliminates the possibility of reuse of analysis modules to ensure the translation to multiple targets.

5.2 Technical suggestions

NooJ is based on the use of transducers describing local grammars. These grammars are used to characterize and resolve linguistic phenomena and some ambiguities. Using transducers, dictionaries and files provided by NooJ, we define an approach based on local grammars to facilitate the applicability of different linguistic methods. The following figure illustrates the proposed approach:

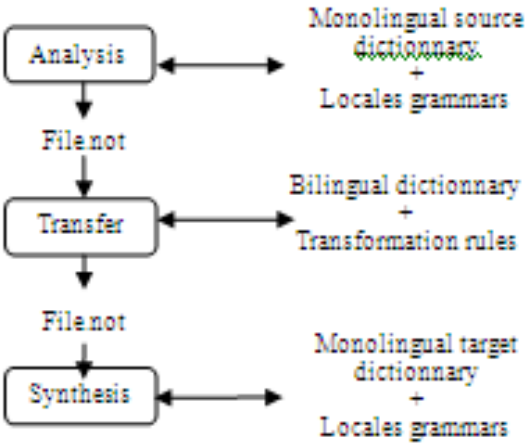


Figure 4. Approach based on local grammars

Figure 4 illustrates the approach based on local grammars that we propose. We use a monolingual dictionary source and local grammars to analyze the structure to translate. During the analysis phase, we seek to resolve linguistic problems and to make appear information for the next phase. The concordance result of analysis is stored in a file extension ".not" to be used as input for the transfer phase. This is performed by using the command "Export concordance". The version analyzed, including the linguistic facts resolved, is transferred to the target lexical space through a bilingual dictionary. This transfer is also provided using transformation rules that adjust certain properties of entities from SL in TL standards. The result of the analysis phase combined with the result of transfer is the input of the synthesis phase. This phase uses local grammars and monolingual dictionaries describing the TL to generate a structure translated correctly.Indeed, these agreements ensure the conjugation of verbs, the endings of words etc. However, the realization of our approach seems possible, our experiments performed showed that some obstacles must be treated. We cite the difficulty of gathering analytic and transfer information and the ability to have concordance that contains external indication.
The ability of NooJ to manipulate files allows us to define a second approach based on XML files that we illustrate by the following figure:

Figure 5. Approach based on the XML formalism

Figure 5 describes the approach based on the XML formalism. The translation process described by the process begins with an analysis phase. The latter provides a syntactic tree describing the source structure to translate. The tree is mapped into an XML file which constitutes the input for the transfer module. After changing the structure stored in XML

file, another file describing in the same type describe the TL is extracted. This can be performed by an external program. From thetarget XML file, a target syntactic tree is derived and the corresponding structure is generated. Thus, the mapping between two trees, each of which describes a language, shows a type of transfer named multilevel. But, we must indicate the incapability of NooJ to extract a tree from an XML file and the difficulty of performing synthesis by NooJ.

6 Conclusions

In this paper, we have presented an overview about linguistic AT. We have tried to resolve problems relating on choice of the suitable method of AT. That's why, we have performed a comparative study between linguistic methods of AT and we have defined a number of criteria facilitating the choice of the appropriate method. In order to deepen our study we use the linguistic platform NooJ. By using NooJ, we have evaluated its capacities in the AT field when testing the applicability of different linguistic methods. In fact, we note the performance of NooJ in term of analysis. But, we have to indicate the difficulties to ensure transfer and synthesis modules. For this, we have presented some suggestions to ameliorate the translation performed. The two main approaches suggested facilitate the realization of transfer and pivot based method.
As perspectives, we note that our comparative study can be extended by a number of points to make the comparison richer. Also, the addition of more capabilities to NooJ makes it more performing concerning AT. These additions allow the extension of AT works using NooJ and give us as researchers a variety of subjects to be treated.

Keywords: AT, linguistic approach, linguistic method, comparative study, selection criteria, NooJ

References

Babych B, Hartley A, Sharoff S. 2007.*Translation from under-resourced languages: comparing direct transfer against pivot translation,* Proceedings of Machine Translation Summit XII, p 68-74, September, Denmark.

Bairrero A. 2008. *port4NooJ: an open source , ontology driven Portuguese linguistic system with applications In machine translation*, acts ofconference NooJ, p 19-47, June, Budapest.

Claveau V. 2007. *Traduction automatique de termes biomédicaus pour la recherche d'information interlingue*, Proceeding of CORIA conference, p 303-318, mars, France.

Dorr B.J, Hovy E.H, Levin L.S. 2004. *Machine translation: interlingual methods, Natural language processing and machine translation, Encyclopedia of language and Linguistics*, 2nd edition. (ELL2), p 1-20.

Fehri H, Haddar K, Ben Hamadou A. 2009.*Integration of a transliteration process into an automatic translation system for named entities from Arabic to French,* Proceeding of NooJ conference, p. 285-300, June, Tozeur, Tunisia.

Papadopoulou E, Gavriilidou Z. 2009. Towards a Greek-Spanish NooJ module, *Proceeding of 2009 NooJ conference*, p. 301-312, June, Tozeur, Tunisia

Sahnoun H, Haddar K. 2009. Comparative study between linguistic methods of machine translation and their experimentation with NooJ, *Proceedings of NooJ Conference*, p. 257-270, June, Tozeur, Tunisia

Silberztein M. 2004. *NooJ: A Cooperative, Object- Oriented Architecture for NLP. INTEX pour la Linguistique et le traitement automatique des langues*, Presses, Universitaires de Franche-Comté, Cahiers de la MSH Ledoux, p 359-370, Besançon, France.

Silberztein M. *2009.* Syntactic parsing with NooJ, *Proceedings of NooJ Conference*, p. 177-189, June, Tozeur, Tunisia.

Wehrli E, Nerima L. *2008.* Traduction multilingue: le projet MulTra, *Proceedings of TALN Conference*, p47-54, June, France.

Proposal of a framework for the representation of Arabic named entities to use the transfer approach with NooJ

Hela Fehri [(1)], Kais Haddar [(2)] and Abdelmajid Ben Hamadou [(3)]

[(1)] *MIRACL-University of Franche-Comte and university of Sfax, Tunisia*
hela.fehri @ fss.rnu.tn
[(2)] *MIRACL-University of Sfax, Tunisia*
kais.haddar @ fss.rnu.tn
[(3)]*MIRACL-University of Sfax, Tunisia*
Abdelmajid.benhamadou @ isimsf.rnu.tn

Abstract

Recognition and translation of named entities (NE) are two current research topics with regard to the proliferation of electronic documents exchanged through the Internet. So, the need for assimilating these documents through NLP tools has become necessary and interesting. Moreover, the formal or semi-formal modeling of these NEs may intervene in both processes of recognition and translation. Indeed, the modeling makes more reliable the constitution of linguistic resources, limits the impact of linguistic specificities and facilitates transformations from one representation to another. In this context, we propose a framework for the representation of Arabic Named entities to use the transfer approach.This framework is based on the structure of features independently of lexical categories.

1 Introduction

The formal or semi-formal modeling of named entities is involved in many fields of information processing. It enables the constitution of linguistic resources to be more reliable. Indeed, such a modeling can represent all the constituents of a named entity in a standard manner and limit the impact of linguistic specificities.

In fact, a formal representation of Arabic named entities (NE) can help, firstly, in the identification of the dictionaries and grammars required for a given application and, secondly, in the use of advanced linguistic methods of translation (i.e., transfer or pivot method). This level of abstraction favors the reuse of certain linguistic resources.

The elaboration of a formal and generic representation of an NE is not an easy task because, on the one hand, we have to find a representation that takes into consideration the concept of recursion and length of NE. In fact, an NE can be formed by other NEs. So, its length is not known in advance. On the other hand, the representation to be proposed should also contain a sufficient number of features that can represent any NE independently of the domain and grammatical category. In other words, the same features must satisfy all types of NE.

It is in this context that the present work is situated. In fact, the main objective is to propose a framework of NE representation to implement the process of recognition and translation of NEs whatever the domain and the chosen type of the domain hierarchy are.

To reach our objective, we have to choose a structure of a framework that takes into account the notion of recursion of NE. We have also to specify features that describe any NE. Finally, the representation should be compatible with the linguistic platform NooJ.

In this paper, we present, firstly, a brief overview of the state-of the art. Then, we describe our proposed framework. After that, we give a general idea of implementation. Finally, the paper terminates with a conclusion and some perspectives.

2 Related work

Research on NEs revolves around two complementary axes: the first involves the typing of NEs while the second concerns the identification and translation of NEs.

As for the identification, tagging and translation of NEs, they were implemented for multiple languages based on different approaches: linguistic (Coates-Stephens, 1993), statistic (Borthwick et al., 1998) and hybrid (Mikheev et al., 1998). Regarding the recognition of NEs, we cite the work presented in (Friburger, 2002). This work allows the extraction of proper names in French. The proposed method is based on multiple syntactic transformations and some priorities that are implemented with transducers. We can cite also the work described in (Mesfar, 2007). The elaborated method is applied on a biomedical domain. Other Arabic works are dealing with the recognition of elliptical expressions (Hasni et al, 2009), with compound nouns (Khalfallah et al, 2009) and with broken plurals (Ellouze et al, 2009).

Other works have been dedicated to the translation of NEs from one language to another. We can cite the work presented in (Barreiro, 2008) dealing with the translation of simple sentences from English to Portuguese. Besides, the work of (Wu, 2008) provides a noun translation of French into Chinese. Her prototype tests a limited corpus of 600 French nouns.

The literature review shows that the already proposed translation approaches are not well formalized (e.g., lack of abstraction and genre). Each one addresses a particular phenomenon without taking into account other phenomena. But, we cannot deny the existence of the formalisms involved to generalize phrase structure grammar (Pollard et al., 1994) or to extract and formalize terminologies (Bourigault, 2002).

Furthermore, all translations using NooJ platform adopt a semi-direct approach of translation, in which the recognition task is combined with that of translation. Thus, the reuse of such works has become limited, which does not promote multilingualism.

3 Proposed framework for representing Arabic NEs

The framework that we propose is used to formalize and to identify Arabic NEs. This framework is inspired from formalisms based on structural features like Head-driven Phrase Structure Grammar (Pollard et al., 1994). Its features are inspired from the concepts "Head and Expansion" introduced by (Bourigault, 2002).

The essential characteristics of the feature structure of the proposed frameworkare:

- the elements of the structure can be atomic or complex,
- the internal structure of an element is defined by its attributes and values.

In what follows, we describe the structure and features of our proposed framework.

3.1 Structure and features of the proposed framework

A NE is composed of two parts: one is essentialand another is extensional. The essential part has a type and remains a NE. Therefore, the latter has itself an extensional and essential part. This proves the recursion for a NE.The type of a NE is usually indicated by a trigger word.The essential part is represented by the feature *"Tête_EN" (head of NE)* and the trigger word is represented by the feature *"Mot_declencheur"*.The extensional partrepresents the final form that composes the NE. It does not admit a type because it is preceded by a lexical item *"Element_EN"(element of NE)*(preposition, special character, …). Then, it can not be considered as a NE but it can contain a NE.Its existence or not does not affect the NE meaning. This part is represented by the feature *"Fin_EN"(end of NE)*.

Therefore, a skeleton of NE structure representation is presented as follows.

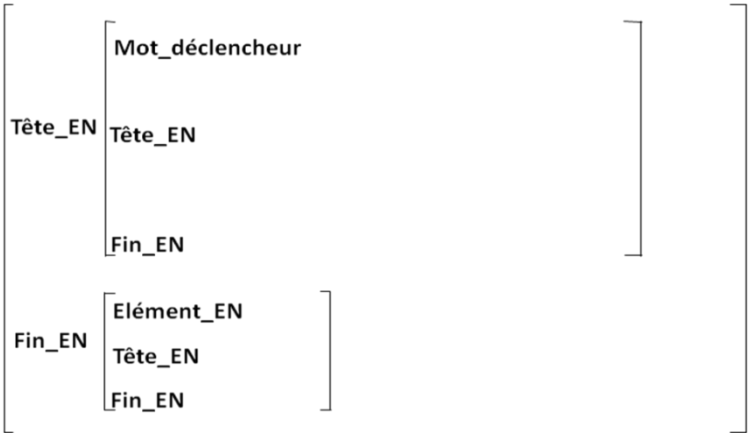


Figure1. Skeleton of NE structure representation

Figure 1 shows that the value of the feature *"Tête_EN"* can be structured or atomic. If it is structured, then it is composed by the features: *"Mot_déclencheur"*, *"Tête_EN"* and *"Fin-EN"*. The *"Mot_déclencheur"* value is simple or composed. Indeed, the trigger word can be formed by a word or a set of words. It can also be empty. The *"Fin_EN"* value can be structured or atomic. If it is structured, then it is composed by the features: *"Element_EN"*, *"Tête_EN"* and *"Fin_EN"*.It can also be empty.

The *"Element_EN"* value is atomic.In the proposed framework, each value of the feature *"Tête_EN"* is typed whatever this value is (structured or elementary). In fact, this feature represents a NE.

3.2 Principles of the proposed framework

Forthe presented framework, two principles should be satisfied and useful in the recognition phase.These principles are used to indicate whether a NE is well formed or not.These principles are the following:

Saturation principle. The structure can be regarded as an NE. Thus, a structure is called saturated if it consists of an NE head whose value is elementary or structured. Figure 2 describes an example of a formal representation that satisfies a saturation principle.

Figure 2. Representation of the word *الرياض*

In Figure2, the value of the feature *"Tête_EN"* is elementary. Thus, a word *الرياضRiadh* is considered a NE whose type is *Ville*.

Non-saturation principle. A structure is called unsaturated if it is formed only by a NE end.

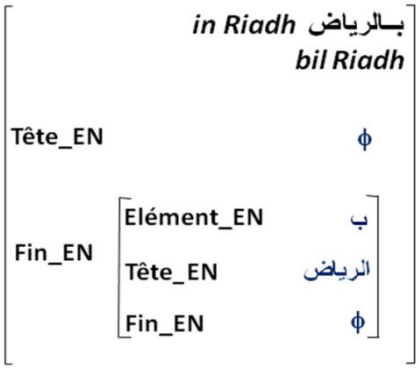


Figure 3. Representation of the word *بالرياض*

In Figure 3, the value of the feature *"Tête_EN"* of the word *بالرياض* is empty. Then, this word cannot be considered as a NE. It doesn't satisfy the saturation principle. But, it should be noted that this word can contain a NE. The two mentioned principles allows as to avoid ambiguity between a word (or set of word) considered as a NE or not.

3.3 Illustrative example

In this section we give an example that explains how constructing NE representation. So, Figure 4 gives a formal representation of the NE ملعب الملك عبد العزيزالدولي بالرياض*Malaab el malik Abd el Aziz el doali bil Riadh*. *King Abd el Aziz international stadium in Riadh.* This NE is described by the following rules:

NE → Trigger word + Function + First name + adjective + Toponym
Trigger word →ملعب*malaab*
Function →الملك*elmalik*

First Name → عبد العزيز *Abd el Aziz*
Adjective → الدولي *el doali*
Toponym → الرياض *elRiadh*

In the NE ملعب الملك عبد العزيز الدولي بالرياض *Malaab el malik Abd el Aziz el doali bil RiadhKing Abd el Aziz international stadium in Riadh*, the word الرياض has the function of place complement. It comes just to elaborate on the NE. Consequently, its elimination dosen't affects the meaning of this NE. That's why; it will be presented by the feature *"Fin_EN"* as follows: *"Elément_EN"* equals ب and *"Tête_EN"* equals *الرياض Riadh*. Let's note the word الرياض *Riadh* has its proper type without the particle ب *bi*. For that reason, it will be a simple value of the feature *"Tête_EN"*. However, the rest of the NE ملعب الملك عبد العزيز الدولي *Malaab el malik Abd el Aziz el doali* has in turn its proper type. It will hence be described by the feature *"Tête_EN"* which will be structured because this NE contains supplementary words. In our NE, the supplementary word is the adjective *الدولي el doali* which has the same role as the place complement بالرياض *bil Riadh*. Therefore, it will be put in the *"Fin_EN"*. Moreover, the word *ملعب malaab* plays the role of trigger word. What is left is the NE *الملكعبدالعزيز el malik Abd el Aziz* whose type is person name. In this NE, the word الملك is also plays the role of trigger word. But, the NE: *عبدالعزيز Abd el Aziz* has a type of first name. So, it will be represented by the feature *"Tête_EN"* as a simple value since *عبدالعزيز Abd el Aziz* doesn't contain any supplement word.

Therefore, the mentioned NE is represented in our framework as follow:

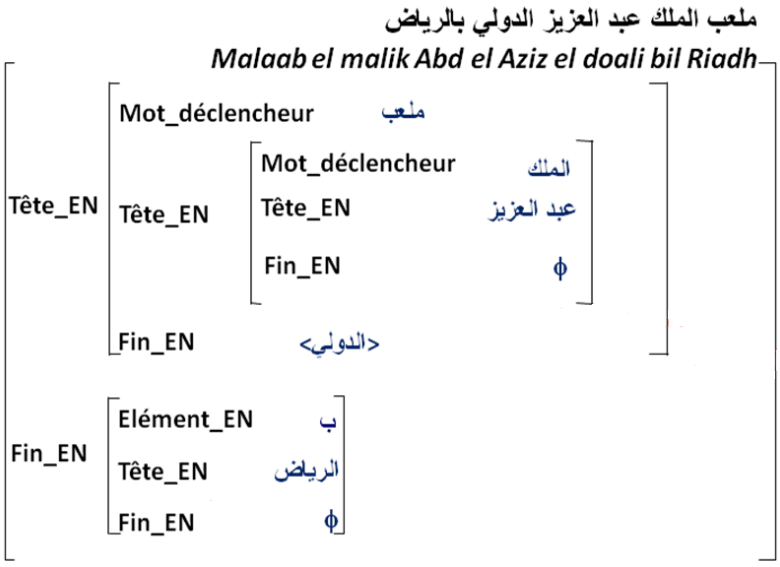


Figure 4. Example of a formal representation

In Figure 4, the saturation principle is determined. In fact, the value of the *"Tête_NE"* feature is not empty.

Let's note that the proposed representation is applicable independently of the domain. In fact, having conducted a study of the location names, we noticed that all their NEs have the same structure whatever the domain is. For example, *the international stadium of king Abdelaziz in Ryadh* has the same structure of *the hospital of Doctor Mohamed Hosni in Ryadh*. So, the trigger word *stadium* is replaced by *hospital*, the trigger word *king* is replaced by *Doctor* and the head NE *Abdelaziz* is replaced by *Mohamed Hosni*. And, the *"Fin_EN"* doesn't change its value. The *"Tête_EN" international* is replaced by empty set and the rest of NEs retain the same value.

3.4 Literal translation representation

Aftercompleting the framework for each NE, we are having a literal translation (word to word) of each feature value composing our proposed framework. Figure 5 represents an example of this phase.

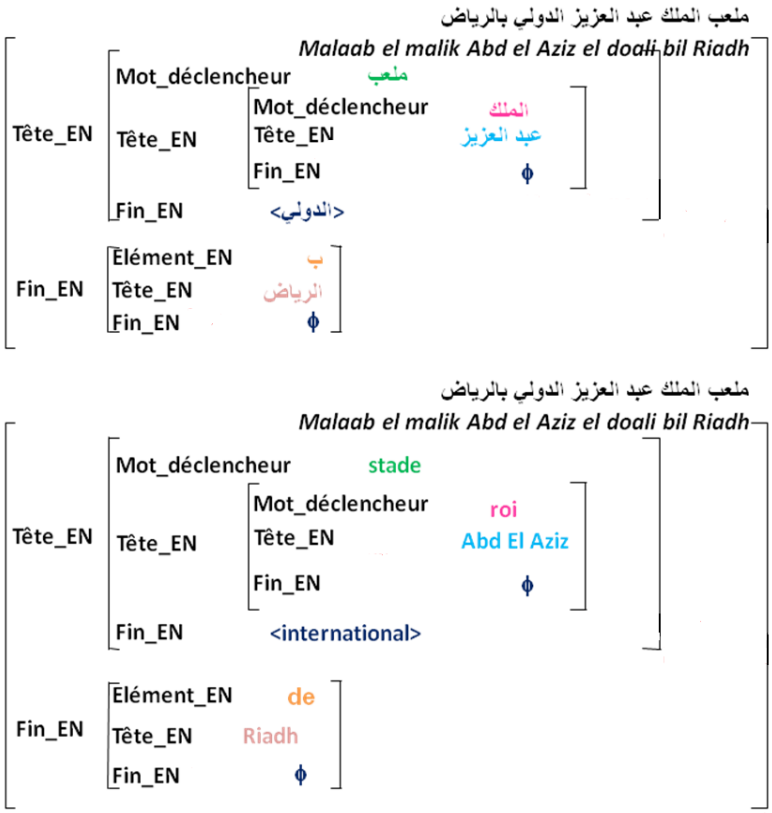


Figure 5. *Example of a literal translation*

As shown in Figure 5, the word ملعب*malaab* is tranltated to *stade*, the word الملك*elmalik* to *roi* and the adjective الدولي*el doali* to *international.*

Several readjustment rules can be deduced at this stage. For example, on the one hand, if a NE in the source language contains an adjective then we have to know whether this adjective belongs to the trigger word or to the noun that comes just before.

On the other hand, if a NE in the source language contains a noun then some rules are applied to solve the problem of contracted forms in Arabic.

4 Contribution of the proposed framework in our approach

From this framework, we can identify the necessary resources for the recognition and translation of NEs. In fact, each structured NE (value *of tête_EN* feature is structured) is transformed into a grammar (nom de lieu *place name*, nom de personne *person name* ...) whereas, each elementary NE (value of *tête_EN* feature is elementary) will be transformed into a dictionary (nom de ville *city name* ...). What is also worth noting is that if we find in the same representation two NEs with the same type, then this indicates that there are two different paths to the recognition of NE (cases of grammar of nom de lieu *place names* in Figure 2).

From the NE representation in the considered framework, we have created the following grammar:

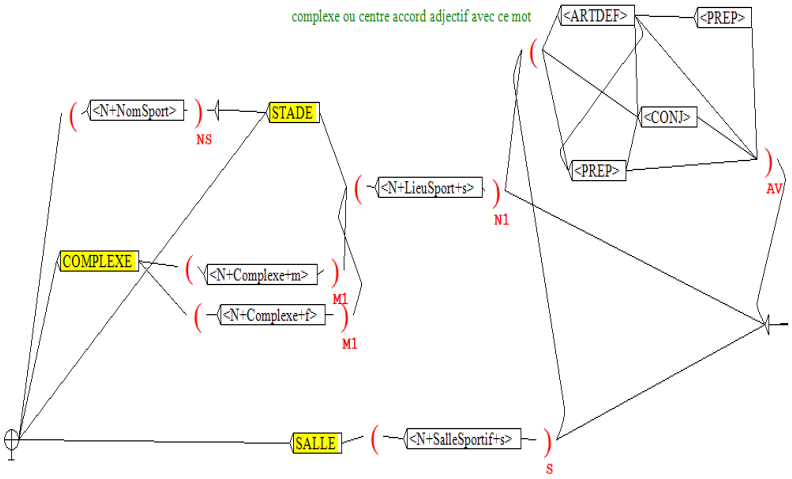


Figure 6. Example of a grammar of recognition of NE

Each path in the grammar of Figure 6 represented a rule extracted in the study corpus. This grammar allows NE recognition.

To implement literal translation phase, we have created the following grammar:

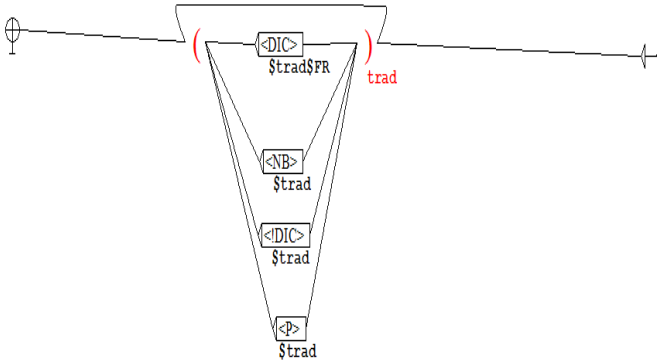


Figure 7.Grammar of literal translation

Let's note that in the grammar of Figure 7, we take into account the special characters that should keep the same value in the target language.

5 Implementation and evaluation

To experiment and evaluate our work, we have applied our approach to two different domains: sport and university institution. We started with the sport since it is the subject of our study corpus. Therefore, we obtained the results given in the concordance table of Figure 7.

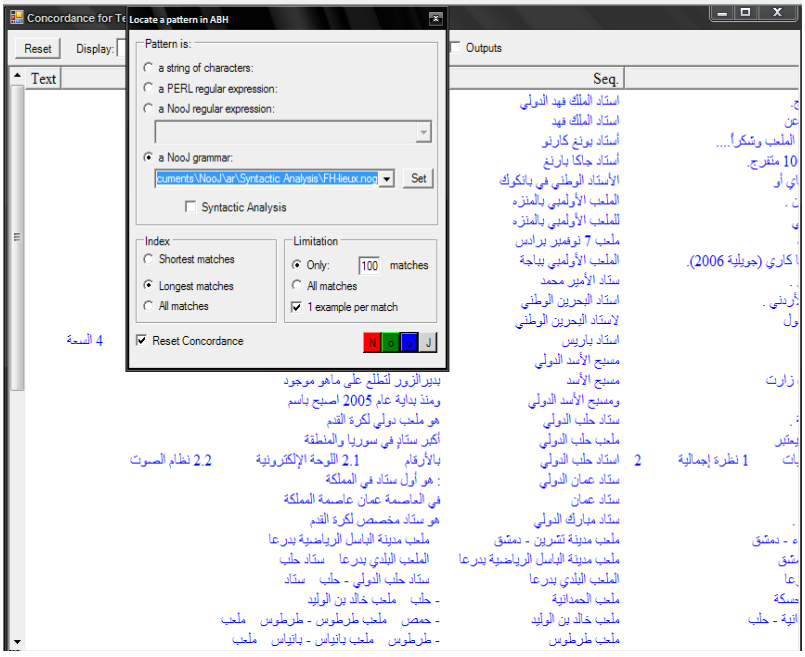


Figure 7. Concordance table of the sport domain

As shown in Figure 7, the concordance table contains different extracted NEs like *الملعبالأولمبيبالمنزه Olympic tadium el-Manzeh, ستادحلبالدولي Alep internationalstadium*.

We have also applied our approach on the university institution domain and we have found the following concordance table.

Figure 8. Concordance table of the university institution domain

So our approach is applicable regardless of the domain provided that we use the same features adopted in dictionaries we have built. Figure 9 gives an idea about these features.

Entry	Example	Annotation in dictionary
Personality Name	الحبيب بورقيبة	N+Perso
Function Name	الأمير	N+Fonction
Toponym	تونس	N+Toponyme

Figure 9. Examples of features used in dictionaries

As shown in figure 9, the personality name is represented by the feature *N+Perso* in our dictionary. To evaluate our result, we used a corpus containing 4000 texts and we obtained 76% of Precision, 84% of recall and 80% of F-measure.

	Precisison	Recall	F-measure
Newspaper texts (domain of sports)	98%	80%	89%

4000 texts

The silence can be explained through the fact that we have some NEs which contain a trigger word followed by a noun without a specific type. For example,the NE *malaab riaaya*cannot be recognized because if it is recognized then it will cause noise such as *The stadium of the next city*.
The proposed representation model is very simple to implement in NooJ. It facilitates the transformation from the semi-direct translation to transfer translation. As well as, it helps the promotion to the reuse of the needed grammars. In fact, it is sufficient to change the inputs (eg, dictionaries, morphological grammars) of the syntactic grammars for the desired results.

6 Conclusion and perspectives

In this paper, we have proposed the proposed a formal framework for representing Arabic NEs eventually NE from other language. We have given an idea of its structure, its features and principles that it should be satisfied. We have also given an experimentation and evaluation on the sports and university institution domains proving that our resources can be reused independently of the domain. The experimentation and the evaluation are done in the linguistic platform

As perspectives, we seek on the one hand, to improve the framework by introducing other features and detailing principles. On the other hand, we seek to finish the implementation of the third phase to achieve the transfer translation.

Keywords: Arabic NE formal representation, NE-translation, Trigger word, Transfer method, Transducers.

References

Barreiro A. 2008. « Port4NooJ: an open source, ontology-driven Portuguese linguistic system with applications in machine translation », *NooJ'08*, Budapest.

Borthwick, A., Sterling, J., Agichtein, E. et Grishman, R. 1998. *NYU:* Description of the MENE Named Entity System as used in MUC-7, in *Proceedings of the Seventh Message Understanding Conference* (MUC-7).

Bourigault D. 2002. UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *TALN*.

Coates-Stephens, S. 1993. The Analysis and Acquisition of Proper Names for the Understanding of Free Text, in *Computers and the Humanities*, Kluwer Academic Publishers, Vol. 26(5-6), Hingham, MA, p. 441-456.

Ellouze S., Haddar K., Abdelwahed A. 2009. *Etude et analyse du pluriel brisé avec la plateforme NooJ*, Tozeur, Tunisie.

Friburger, N. 2002. *Reconnaissance automatique des noms propres*, PhD thesis, university of François Rabelais.

Hasni E., Haddar K., Abdelwahed A. 2009. Reconnaissance des expressions elliptiques arabes avec NOOJ, in *Proceedings of the 3rd International Conference on Arabic Language Processing* (CITALA'09) sponsored by IEEE Morocco Section, 4-5 May 2009, Rabat, Morocco, pp 83-88.

Khalfallah H. F., Haddar K., Abdelwahed A. 2009. Construction d'un dictionnaire de noms composes en arabe, in Proceedings *of the 3rd International Conference on Arabic*

Language Processing (CITALA'09) sponsored by IEEE Morocco Section, 4-5 May 2009, Rabat, Morocco, pp111-116.

Mesfar S. 2007. *Named Entity Recognition for Arabic Using Syntactic grammars*. NLDB 2007 Paris, 28-38.

Mikheev, A., Grover, C. et Moens, M. 1998. Description of the LTG system used for MUC -7, in *Proceedings of 7th Message Understanding Conference* (MUC-7), http://www.itl.nist.gov/iad/ 894.02/ related_projects/muc/.

Pollard C., Sag., I.A. 1994. *Head-Driven Phrase Structure Grammar*, published by the press in the University of Chicago, Edition Golgoldmittu, Chicago, LSLI.

Wu M. 2008. *La traduction automatique français-chinois pour les groupes nominaux avec Nooj*,Oral Presentation at NooJ 2008 Conference, Budapest.

Greek Professional nouns processed with NooJ

Chadjipapa Elina[(1)], Papadopoulou Lena[(2),]

[(1)] *Democritus University of Thrace, elinaxp@hotmail.com*
[(2)] *Autonomous University of Barcelona, lepapad@hotmail.com*

Abstract

The aim of this paper is to present the processing of Greek professional nouns with NooJ platform. A series of morphological and syntactic grammars have contributed to the construction of the Greek Professional Noun Lexicon (G.P.N.L). Working on the elaboration of the G.P.N.L with NooJ some significant observations of Greek professional nouns have been made and will be further outlined.

1 Introduction

The results of the linguistic analysis of our corpora demonstrated that a significant number of professional nouns were considered unknown or their annotation presented ambiguity problems caused mainly by their polysemous nature. Thus, it was decided to work on the professional nouns with a view to improve the automatic processing of the Modern Greek language with NooJ.

In this paper, the theoretical framework of our work will first be presented. Step by step description of the construction of the G.P.N.L. will follow, outlining the macrostructure and the microstructure of the dictionary as well as the syntactic and morphological grammars that have been built in order to facilitate the elaboration of the G.P.N.L. Finally, interesting observations regarding Greek professional nouns will be stated.

2 Theoretical Framework

Gavriilidou (2006: 145), who has initiated the study of the professional nouns, comments that "professional nouns constitute a well defined *class of objects* within Greek vocabulary".

Classes of objects (Gross, 1992) is the theoretical framework on which our work is principally based on. We will precisely state that *classes of objects* are semantically homogeneous classes of lexical units based on syntactic criteria. In other words, according to this theoretical model for every predicate, information is provided regarding to the semantic type of the arguments. An elementary class of arguments can be defined from only one predicate that could be nominal, adjectival or/and verbal. The predicates can further be defined through the same procedure. Following the order that predicates presented, above we will view them one by one. When the predicate has the basic form of a noun its definition depends on the support verb that selects it. When the predicative form is an adjective its definition is based on the noun that determines it. Finally, when the predicate is a verb (as it is in our case) its definition is related to the class of arguments that selects.

In Greek, the class of <*Professions*> is defined by verbal arguments and selects verbs such as: *επαγγέλλομαι*/EN: work) as appropriate operators and verbal phrases such as: *ασκώ το επάγγελμα του*/EN: - practice a profession). Thus, every noun that appears as complement of these verbs/verbal phrases will be added to the class of <*Professions*>.

Further, the model of classes of objects extends the description of the lexical units, searching for special adjectives for each class. These adjectives characterize the nouns of each homogeneous semantic class. Some of them in the class of <*Professions*> for the Greek language are: *ηλεκτρονικός/τεχνικός/υπεύθυνος/μηχανικός* etc. Finally, collocations and frozen phrases are examined in order to separate the literal from the metaphorical forms. Some examples are quoted: *Δάσκαλε που δίδασκες και λόγο δεν κρατείς/ λογάριαζε χωρίς το ξενοδόχο/ ο καλός ο καπετάνιος στη φουρτούνα φαίνεται/ ένα μήλο την ημέρα τον γιατρό τον κάνει πέρα.*

Thus, the model of *classes of objects* contributes the lexical entries, not as individual words but as phrases that are categorized according to their syntactic characteristics. Hence, a lexicon-grammar is created according to which the lexicon and the grammar are depended to each other.

Apart from the aforementioned theoretical model some considerable lexicographic models for the elaboration of the G.P.N.L followed. First, they were based on the *monolingual coordinated dictionaries* of Blanco (2001), which comprises a series of theories, the *lexicon-grammar*, *classes of objects* and the introduction of the *domains*. The domains proposed by Buvet and Mathieu-Colas (1999) contributed to the construction of the microstructure of the G.P.N.L. (see 3.2). Finally, the DicPro of Fuentes (2005), which concerns a monolingual coordinated dictionary of Spanish and French professionals and the electronic dictionary of French-Spanish-Catalan-Arabic professional nouns of Blanco and Lajmi (2004), were the two lexicons that our study was based on.

3 Collecting data

Modern Greek dictionaries of Triantafyllidis[1] (1998) and Babiniotis (1998) constituted the main sources for the compilation of our dictionary. Apart from these two dictionaries, a series of vocabulary lists, diverse web pages, such as the http://www.ypakp.gr, and the Hellenic corpus, which had been created for the first version of the Greek NooJ Module were supplementary sources for the macrostructure of our dictionary. The morphological and syntactic grammars that we created especially for professionals were also a useful tool for the enhancement of our data.

3.1 Morphological grammar

As we have already mentioned, we created a morphological grammar in order to extend the macrostructure of our dictionary with professional nouns in a semi-automatic way. This grammar allowed us to annotate Greek professionals encountered in our corpus automatically and, thereafter, to introduce them to the macrostructure of our dictionary, where they have been further lexicographically treated.

The construction of the morphological grammar (Figure 1) is based on the suffixes (*-ίστας/ γραφίστας*/EN: commercial artist), confixes (*-λόγος/ανθρωπολόγος*/EN: anthropologist) and

[1] We used the electronic version which is available in the web site www.komvos.edu

productive roots (-τεχνίτης/ *οδοντοτεχνίτης*/EN: dental technician) that Greek professional nouns usually pose (Triantafyllidis, 1996 Anastasiadis-Symeonidis, 2002). According to this grammar every letters sequence (<W>) ending to –ωρύχος, -τρόφος, -τεχνίτης etc will automatically be annotated as a human professional noun.

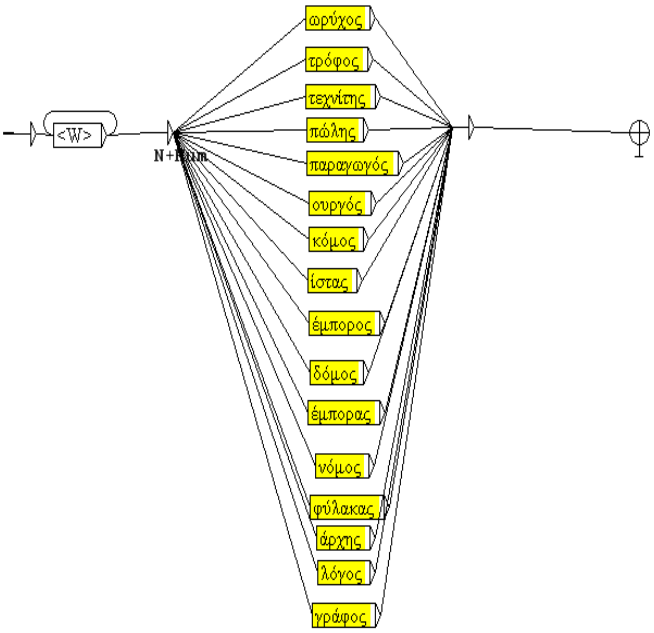


Figure 1. Morphological grammar for Greek professional nouns

As it is known Greek is a heavily inflected language. This means that the creation of a morphological grammar that simply quotes the suffixes is not enough. In other words, we had to provide to NooJ in separate graphs (Figure 2) all the inflective forms of the suffixes in order all the word forms of the professionals to be annotated.

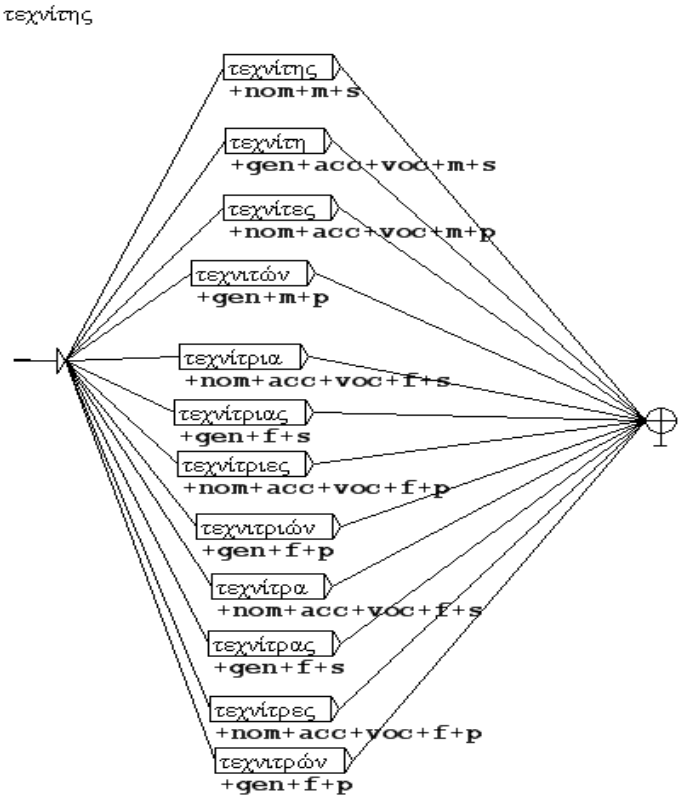


Figure 2. Graph of the inflexional paradigm of the productive root -τεχνίτης

However, before applying the morphological grammar, we had to test if it annotates the professional nouns automatically and register the exceptions of this grammar. In other words, we had to clarify problems homonymy. That means that we encountered words ending to the sixteen aforementioned cases but were not professions, as we can see to the following examples:

- Productive lexical roots: -*φύλακας*
  (e.g. χαρτοφύλακας, EN: briefcase/ αστυφύλακας, EN: police officer)
- Confixes: - γράφος
  (e.g. λεξικογράφος, EN: lexicographer/ φασματογράφος, EN: spectrograph)
- Suffixes : - ίστας
  (e.g. γραφίστας, EN: commercial artist/ τουρίστας, EN: tourist)

The registration of these exceptions in the *Show Contract* application of NooJ, debated the emerged problem of homonymy. However, we do not pretend that the grammar can cover the totality of the cases, as we have to deal with a natural language

3.2 Syntactic grammar

Apart from the morphological grammar we also built a syntactic grammar in order to pick up automatically the professionals of our texts and introduce them to our macrostructure. This grammar according to the theory of *class of objects* constructed following syntactic criteria. As every class is associated with a minimal syntactic definition, the class of <*Professions*> is defined from verbal arguments and presents five appropriate operators:

a. *επαγγέλλομαι*/(EN: occupy),
Ο Θάνος επαγγέλλεται τον *(αρχαιολόγος + γραμματέας + δημόσιος υπάλληλος + κομμωτής). Thanos occupy the (archeologist + secretary + civil servant + hairdresser)*

b. *εργάζομαι*/(EN:work), *δουλεύω*/(EN: work), *απασχολούμαι*/(EN: employ), *διορίζομαι*/(EN: be hired)
Ο Θάνος (δουλεύει + εργάζεται+ απασχολείται ως/σαν (αρχαιολόγος + γραμματέας + δημόσιος υπάλληλος + κομμωτής). Thanos work as (archeologist + secretary + civil servant + hairdresser)

c. *ασκώ το επάγγελμα του*/(EN: - practice the profession of),
Ο Θάνος ασκεί το επάγγελμα του *(αρχαιολόγος + γραμματέας + δημόσιος υπάλληλος + κομμωτής). Thanos practice the profession of (archeologist + secretary + civil servant + hairdresser)*

d. *είμαι[…] στο επάγγελμα.* (EN: be - in profession)
Ο Θάνος είναι (βιολόγος + γλωσσολόγος +δικαστής) στο επάγγελμα
Thanos is (biologist + linguist +judge) in profession.

Below (*Figure 3*) is presented the graph of the syntactic grammar including the five operators:

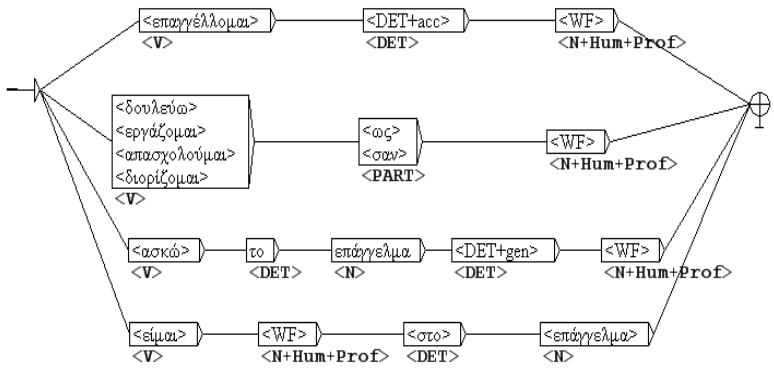


Figure 3. Syntactic grammar for professional nouns.

Continuing, Gavriilidou (2006: 145) mentions, some more syntactic construction such as this with a zero determiner after the verb *είμαι* 'be' can be entered:

Ο Θάνος είναι (αρχαιολόγος + γραμματέας + δημόσιος υπάλληλος + κομμωτής)
Thanos is (archaelogist + secretary + civil servant + hairdresser),

and one with more complicated determiners can be used:

Ο Θάνος είναι (Ε+ο) κομμωτής της μαμάς μου
Thanos is (E [3]+the) hairdresser of my mother'.

Although the last two constructions offer syntactic definition to the class of <*Professions*> which were considered general enough for the specific purpose that we constructed the syntactic grammar and that is the reason why they are not included.

4 G.P.N.L. Construction

After the construction of the grammars and their application to our corpus, the introduction of the professionals to our macrostructure along with the lemmas encountered by the aid of other sources (see 2), their lexicographic treatment in the microstructure of the GPLN followed.

4.1 Macrostructure

The macrostructure of our Lexicon consists of 4.200 lemmas. The 3.000 of them are simple nouns and the 1.200 are compound nouns, which have also been treated as simple words. The compound nouns that are included in our lexicon present various structures, such as:

Noun-Noun: *αρχίατρος στρατού*/EN: chief medical officer
Adjective-Noun: *κοινωνικός λειτουργός*/EN: social worker

Noun-Preposition-Noun: *εργαζόμενος σε ασθενοφόρο*/EN: ambulance worker

4.2 Microstructure

The microstructure consists of seven fields that provide linguistic information for each lemma:

- The **G field** presents the grammatical category of our lemmas. All of them are codified as **N,** as they are all nouns.
- The **F field** includes the codification of the inflectional paradigm to which belong.
- The **O field** contains the opposite gender of each professional noun. We have to mention that there are cases in which there is no opposite gender. For example, there is no feminine of the professional noun *παπάς* (EN: priest), as the feminine *παπαδιά* do not correspond to a professional noun, but it defines the relationship between the priest and his wife.
- The **T field** provides a semantic categorization of each entry; the syntactico-semantic features of our lemmas. In our case only the Human feature has been used, as the professional nouns refer only to Humans.

- The **C field** g*ives m*ore specific semantic information, by providing the classe of object of each lemma. It has to be noted that obviously all lemmas belong to the class of object <*Profession*>.
- The **D field** presents the domain to which each entry belongs, i.e. Theology, Law, Military, Medicine, Economy, Education, Music, sport, religion etc**.**
- The last field is the **EN field.** This field provides the English translation equivalent of each entry.

Thus, the entries of the G.P.N.L present the following structure, as we can see below:

- *γυναικολόγος*,N+FLX=N18+fem=γυναικολόγος+Hum+Prof+Medicine+ EN= gynecologist
- *μεσίτρια χρηματιστήριου*,N+FLX=N27INDEC+mas=μεσίτης χρηματιστήριου+Hum+Prof+Economy+EN= broker agent

5 Some observations

During the elaboration of the G.P.N.L we made some observations concerning the morphology of the Greek professional nouns that are worth mentioning. Starting with the *historical professions, which are professions that do not exist in present days; we have noticed that the great majority* of them ends in particular suffixes, such as:

- - άς e.g. *αβγουλάς*/EN: egg dealer)
- -τζής e.g. *παλιατζής*/EN: scrap dealer)

We also have noticed that most of the female historical professional nouns end at:

- -ου e.g. *μυλωνού*/EN: miller)

In addition, the study of our data allowed us to draw the general conclusion that male professional nouns outnumber the female ones and that there were very few female professional nouns that do not pose a corresponding masculine, such as the professional:

- *νταντά*/EN: nanny)

We should also mention that there is a number of masculine professions that is morphologically the same with the respective female, know as *epicene nouns*.

e.g. m: *πρύτανης*/EN: rector), f: *πρύτανης*/EN: rector)

Although they are morphologically similar, a different inflectional paradigm has been attributed to each case in order to distinguish between the two genders.

e.g. m: *πρύτανης* FLX= N22c /EN: rector, f: πρύτανης FLX= N22h/ EN: rector

Another reason for this distinction was the further use in syntactic grammars, where the gender assignation is considered indispensable.

Finally, deepen in the same morphological phenomenon (epicene nouns) professions end in -ας was examined. According to Ιορδανίδου and Μάντζαρη (2005) the feminization of the male professional nouns ending in -ας (e.g. *m: λέκτορας /EN: lecturer/ f: λέκτορας /EN: lecturer)* was dominated, as these professions existed only in the male gender. Nowadays, they are used in both genders and their feminization caused changes to the

Greek morphological system. As a solution, were suggested various ways of which we partially adopted what Τσοπανάκης (1982) proposed and we applied it to the G.P.N.L. Thus, according to him the epicene type maintains and the feminine gender inflects according to the masculine[2] (e.g. **nom+m+f+s:** *ο/η λέκτορας*, **gen+m+f+s:** *του/της λέκτορα*). Also according to the statistical results of the research of Ιορδανίδου and Μάντζαρη (2005) we introduce a second type for the feminine ending in *–έας* in the genitive of the singular number (e.g. **nom+f+s:** *η διερμηνέας/EN:interpreter* **gen+f+s:** *της διερμηνέα* second type **gen+f+s:** *της διερμηνέως*). Finally, it should be mentioned that there were already existed inflexional paradigms for the masculine professions ending in *–ας* in our previous grammars but not any for the respective feminine. So, we created five new inflexional paradigms from N34a to N34e.

6 Conclusions

In this paper we have presented the Greek Professional Noun Lexicon. A morphological and a syntactic grammar contributed to its construction. Its macrostructure consists of 4.200 lemmas and its microstructure is based on a series of theoretical frameworks.
With the aim to extend the Greek NooJ Module we will try to exhaust the professional nouns by collecting all their respective adjectives, nouns referring to place, instrument, time, product and finally abstract nouns. Moreover, the next step in our work will be the research on further classes of objects such as the class of <clothes>.

References

Αναστασιάδη-Συμεωνίδη, Α., 2002. *Αντίστροφο Λεξικό της Νέας Ελληνικής*, Ινστιτούτο Νεοελληνικών Σπουδών [Ίδρυμα Μανόλη Τριανταφυλλίδη], ΑΠΘ.

Blanco, X. 2001. Dictionnaires électroniques et traduction automatique espagnol-français. *Langages* (143), 49-70.

Blanco, X., Lajmi, D. 2004. "Dictionnaire électronique français-espagnol-catalan-arabe des noms des professions et des métiers", *Actes des Premières Journées Scientifiques des Réseaux de Chercheurs de l'AUF*, Agencia Universitaria de la Francofonía.

Blanco, X. & MEJRI, S. (eds.) 2007. *Los nombres de profesiones: enfoques lingüísticos, contrastivos y aplicados*. Bellaterra: Universitat Autónoma de Barcelona (230 pp.). ISBN 978-84-490-2511-2

Buvet, P.-A., & Mathieu-Colas, M. 1999. Les champs Domaine et Sous-Domaine dans les dictionnaires électroniques. *Cahiers de Lexicologie 75*, pp. 173-191.

Fuentes, S., 2005. Dicpro: dictionnaire électronique des noms de professions, *Workshop Luso Español sobre gramática constrastiva*.

Gavriilidou, Z., 2006. Les noms de professions en *–λόγος*/*-logue* en grec et en français, in (X. Blanco & S. Mejri éds), *Les noms de professions. Approches linguistiques, contrastives et appliquées*, Servei de publicacions, Universidad Autonoma de Barcelona, pp. 128-145.

Gazeau, M. A., Maurel, D., 2006. Un dictionnaire INTEX de noms de professions : quels féminins possibles ? *Cahiers de la MSH Ledoux*. p. 115-127.

[2] Also Τσοπανάκης (1982) proposes an alternative type for the feminine in the genitive of the plural number that we have not adopted.

Τριανταφυλλίδη Μ. 1996. *Νεοελληνική γραμματική (της δημοτικής). Ανατ. της έκδοσης του ΟΕΣΒ (1941) με διορθώσεις*. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Ινστιτούτο Νεοελληνικών Σπουδών, [Ίδρυμα Μανόλη Τριανταφυλλίδη], Θεσσαλονίκη.

Τριανταφυλλίδης, Μ. 1953. «Η "βουλευτίνα" και ο σχηματισμός των θηλυκών επαγγελματικών ουσιαστικών», *Άπαντα Μανόλη Τριανταφυλλίδη*, τόμ. 2, Θεσ/νίκη 1963, σ. 326-334.

Τσοπανάκης, Α 1982. «Ο δρόμος προς τη δημοτική: θεωρητικά, τεχνικά και γλωσσικά προβλήματα. Σχηματισμός των θηλυκών επαγγελματικών», *Νέα Εστία* 1204, σ. 1120-1142.

Dictionaries

Τριανταφυλλίδης, Μ., 1998, *Λεξικό της Κοινής Νεοελληνικής*, Ινστιτούτο Νεοελληνικών Σπουδών, Θεσσαλονίκη.

Μπαμπινιώτης, Γ., 1998, *Λεξικό της Νέας Ελληνικής Γλώσσας,* Κέντρο Λεξικογραφίας, Αθήνα.

Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform

Abdelmajid Ben Hamadou [(1)] Odile Piton [(2)] and Héla Fehri [(3)]

[(1)] MIRACL-University of Sfax, Tunisia.
[(2)] SAMM-University of Paris1 Pantheon-Sorbonne, Paris, France.
[(3)] MIRACL-University of Franche-Comte and University of Sfax, Tunisia

Abstract

The extraction of relation between Named Entities (NE) has become the last few years an interesting research domain. It is very useful for many applications such as Web mining, Information extraction and retrieval, Business intelligence, Automatic databases filing with Entities & types, Questions answering task and document Summarization.
Several works has been performed for relation discovery in texts written in Latin languages and as far as we know, very few works has been done for Arabic language. In this paper, we focus on functional relations between ENAMEX and ORG Arabic Named Entities. The extraction approach is rule based and the implementation is performed using NooJ Platform.

1 Introduction

The extraction of semantic relations between Named Entities (NE) has become the last few years an interesting research domain. It is very useful for many applications such as Web mining, Information extraction and retrieval, Automatic databases filling with Entities & types, Questions answering task and document Summarization.

Relations between Named Entities can be binary (involving two entities) or more complex up to notion of event. There are also several types of relations based on the types of the involved Named Entities.

Several works related to semantic relations discovery has been performed for European languages and especially for English. As far as we know, very few works has been done for Arabic language.

In this paper we propose a rule based relation extraction system for Arabic language. We focus on functional relations between ENAMEX and ORG Named Entities (director, responsible, president…). Functional Relation can be explicit or implicit. Explicit relation is indicated by a special word or sequence of words in the text. Implicit relation is a relation that can be mined from the text using the context.

The extraction process is performed in three steps: Identification of the Named Entities (ENAMED and ORG), Detection of relations between identified Named Entities, Generation of the predicate form representing the relation in Arabic and French. The translation of the relation allows cross lingual information retrieval.

The rest of the paper is organized as follows. We begin by giving an idea about the proposed approaches for extraction of semantic relations between NEs. Then, we address the major challenges posed by extraction of semantic relations between NEs for Arabic language. After that, we detail our approach and its implementation using NooJ linguistic

platform [Silberstein 2005]. Finally, we present the evaluation results obtained from a test corpus.

2 Approaches for NE relation Discovery

Discovering Relations between Named Entities can be done using a linguistic or a numerical or hybrid approach.

The first approach is rule-based, which tried to use syntactic and semantic patterns to capture the corresponding relations by means of manually written linguistic rules. This approach is very interesting for restricted domain and has a good quality of analysis. The major drawback of this approach is the poor adaptability and the poor robustness in handling large-scale or new domain data. This is due to two reasons: rules have to be rewritten for different tasks or when the application is enlarged to different domains and generating rules manually is quite laborious and time consuming (Santos and al., 2010).

The statistical approach (Jun and al., 2009), (Culotta and al., 2006) is based on a learning process from an annotated corpus which can be supervised when the corpus is large, or weakly supervised when the corpus is reduced, or unsupervised if the corpus is not needed. The supervised approach (Miller S. and al., 2000), (Zelenko and al., 2003), (Culotta and al., 2004) and (Kambhatla and al., 2004) costs time and efforts to annotate the corpus, and its performance depends on the size of the corpus. So, to decrease the corpus annotation requirement, some researchers turned to weakly supervised learning approaches, which rely on a small set of initial seed (bootstrap) instead of a large annotated corpus (Agichtein and al., 2000), (Stevenson, 2004). However, the major problem lies in selecting the initial corpus and deciding its "optimal" size.

The unsupervised approach works effectively on high-frequent entity pairs. However, this approach has a limited quality in analysis (Hasegawa and al., 2004).

3 Arabic Named Entities Relations discovery challenges

Arabic Named Entities Relations discovery is not an easy task. In fact, besides the problems related to the Arabic Named Entities recognition (Ben Hamadou and al., 2010), (Shaalan and al. 2009), the extraction of relations between Named Entities poses some specific challenges:

- Discontinuity of the multiple relations concerning the same NE (person).

أ.د. عماد أبو الرب الأمين العام لجمعية كليات الحاسبات - عميد كلية العلوم *Prof. Dr. Imed Abou-Roub the secretary general of organization of computational faculties – Dean of Science Faculty*

The Person Name: *Prof. Dr. Imad Abou-Roub* is concerned by the relation *Secretary-General* and the relation *Dean-of* in the same time. Between the Person Name and the second relation (*Dean-of*) there is a gap.

- Implicit Relations: they are relations that are not directly specified in the text. They are mined from the text using contextual elements.

الدكتور عبد الهادي موسى، اللجنة الشعبية لكلية الطب *Doctor Abd el Hadi Moussa, pupil's committee of medicine Faculty*

The relation between the two NEs: *Doctor Abd el Hadi Moussa* and *pupil's committee of medicine Faculty* is not explicitly indicated by a specific word in the text. In this example the affected relation is *Belong-to*.

- Necessity to use the previous context of the relation in order to know the missing element involved in the relation.
  الأستاذ الدكتور صالح هاشم الأمين العام*Professor Doctor Salah Hachem the secretary general*
  In this example the ORG named entity is absent, but it can be recovered from the right context.

- Interference between implicit relation and discontinuity. In the text we have in the same time implicit relation and discontinuity as shown in this example.

أ.د. القاسم علي القاسم - كلية الزراعة – جامعة الخرطوم / مساعد الأمين لجمعية كليات الزراعة

Prof. Dr. El-Kacem Ali El-Kacem – Agricultural Faculty – university of Khartoum / assistant of secretary of faculties' agriculture organization.

The first implicit relation is*Belong-to* and there is a discontinuity between the Person Name *Prof. Dr El-Kacem Ali El-Kacem* andthe Relation *Assistant-of.*

4 The proposed approach

As indicated above, the proposed approach for relation discovery is rule based. It is founded on a balance between grammar and lexical resources. The grammar indicates the composition rules of the lexical components, in order to form the different patterns of functional relations. Patterns are transformed into transducers implemented using NooJ Platform.

The lexical resources correspond to one dictionary for each entity type used by the transducers.

a. Patterns of relations

The approach of extraction functional relations between ORG and ENAMEX Named Entities is based on a notion of linguistic pattern that we transform into rules and transducers.

Patterns are considered as regular expressions integrating different elements representing relations and concerned Named Entities.

These patterns are identified semi-automatically from our journalistic learning corpus using NooJ facilities. Indeed, NooJ allows to identify regular expressions in a corpus such as all First Names with the right and the left contexts, all passengers containing a specific relation with the right and the left context.

In the following, we give a list of the main patterns identified in the learning corpus:

<Pattern 1>:= {<Title>} <PersName> {<P>} <REL>< ORG >
المهندس علي التويجري مدير المجمع الكميائي
Engineer Ali Al-Touijri Director of Chemical group
<Title>***Engineer*** **المهندس**
<PersName>***Ali Al-Touijri*** **علي التويجري**
<REL>***Director-of*** **مدير**
<ORG>***Chemical group*** **المجمع الكميائي**

<Pattern 2>:= <REL><ORG > {<P>} {<Title>} <PersName>
المهندس علي التويجري مدير المجمع الكميائي:
Engineer Ali Al-Touijri Director of Chemical group
<Title>***Engineer*** **المهندس**
<PersName>***Ali Al-Touijri*** **علي التويجري**
<REL>***Director-of*** **مدير**
<ORG>***Chemical group*** **المجمع الكميائي**
<P>:

<Pattern 3>:= {<Title>} <PersName> {<P>} <ORG >
المهندس علي التويجري/ المجمع الكميائي
Engineer Ali Al-Touijri Director of Chemical group
<Title>***Engineer*** **المهندس**
<PersName>***Ali Al-Touijri*** **علي التويجري**
<Implicit-REL>***Belong-To***
<ORG>***Chemical group*** **المجمع الكميائي**
<P>/

<Patern 4>:= <REL>< ***demonym*** -ADJ> {<Title>} <PersName>
الرئيس الأمريكي : باراك أوباما***US President : Barack Obama***
<REL>***President*** **الرئيس**
< ***demonym*** -ADJ>***American*** **الأمريكي**
<PersName>***Barack Obama*** **باراك أوباما**
<P>**:**

<Patern 5>:= <REL><Toponym > {<Title>} <PersName>
رئيس الولايات المتحدة : باراك أوباما***US President : Barack Obama***
<REL>***President*** **رئيس**
<Toponym >***US*** **الولايات المتحدة**
<PersName>***Barack Obama*** **باراك أوباما**
<P>**:**

{ V } Means that the category "V" is optional

<P> Means any punctuation: , /, (-…

b. NooJ grammar

The figure 1 shows the main graph allowing the recognizing of the functional relation between ENAMEX and ORG Named Entities. Each path of this graph represents a different pattern.

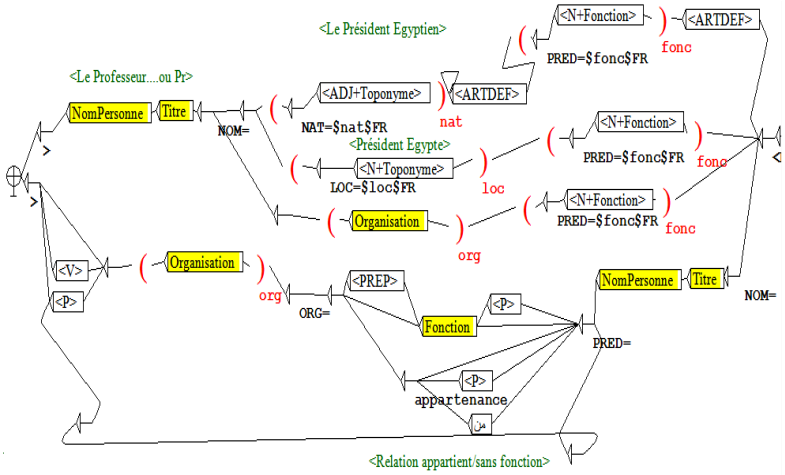


Figure 1. The main graph

The sub graphs:

- *NomPersonne* allows to recognize the Person Names
- *Titre* allows to recognize triggers words preceding the Person Name
- *Organisation* allows the recognition of the ORG Named Entity.
- *Fonction* allows the recognition of the relation between Person Name and ORG Named Entities.

The figure 2: details the *Organisation* graph. It allows the recognition of the ORG Named Entity witch has a complex structure and it includes ENAMEX (Person Name) and LOC (i.e. SN Lieu) Named entities.

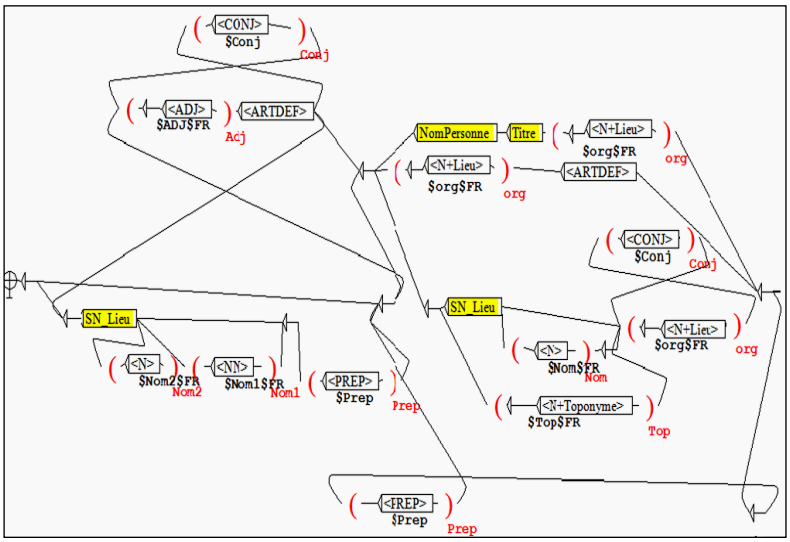


Figure 2. Organisationgraph

Beside those graphs, we use a graph for splitting the agglutinated particles (prepositions, definite articles, conjunctions,..) in the analysed Arabic words as shown in this example :
Analysed word: وبالإدارة *et dans l'administration*
Result: Conjonction = و + Préposition = ب +Définite Article = + ال Noun = إدارة

c. NooJ Dictionaries

As mentioned above, the recognition of the relations is based on the following main modules:

- Recognition of the Person Names, including the possible Title,
- Recognition of the ORG Named Entity
- Recognition of the Relation between the indicated Named Entities.

Each module uses its own dictionaries whose entries contain lemmas with the corresponding possible flexional model to generate all derived forms and the corresponding French lemmas.

The recognition of the Person Names module uses the following dictionaries:

- For the Person Names Recognition Step, we build the following dictionaries:
  - Titles Dictionary
  - First Names Dictionary
  - Last Names Dictionary
- For the ORG Recognition module, we build the following dictionaries:

- Geographical names Dictionary
- Type institution Dictionary
- Adjectives Dictionary

- For the Recognition of relations module, we build the following dictionaries:
  - Demonym adjectives Dictionary
  - Functions Dictionary

Extracts of Arabic Dictionaries
 # Title/Functions names
مدير,N+Fonction+FLX=A1+FR=directeur
ملك,N+Fonction+FLX=ملك+FR=roi
أمين,N+Fonction+FLX=رئيس+FR="Secrétaire"
نائب,N+Fonction+FLX=A1+FR="vice"
مهندس,N+Fonction+Titre+FLX=A1+FR="Ingénieur"
 # names of geographical categories
جمهورية,N+FLX=قارة+Cat_Geo+Toponyme+FR=république
مملكة,N+FLX=قارة+Cat_Geo+Toponyme+FR=royaume
 # Geographical names
تونس,N+PR+s+Pays+Toponyme+FR=Tunisie
تونس,N+PR+s+Ville+Toponyme+FR=Tunis
رياض,N+PR+s+Ville+Toponyme+FR=Riyadh
 # Personalities' names
htebasilE=RF+s+f+osreP+RP+N,تابيع لبزبا
حبيب بورقيبة,N+PR+Perso+m+s+FR="Habib Bourguiba"
 # adjectives
lanoitan=RF+1A=XLF+A,وطني
أولمبي,A+FLX=A1+FR=olympique
بلدي,A+FLX=A1+FR=municipal
lanoitanretni=RF+1A=XLF+A,دولي
 # Demonym adjectives
neisinut=RF+1A=XLF+emynopoT+A,تونسي
neitpygé=RF+1A=XLF+emynopoT+A,مصري
 # Institutions
إتحاد,N+Lieu+FR=union
جمعية,N+Lieu+FR=association
مجمع,N+Lieu+FR=confidération
كلية,N+Lieu+FLX=قارة+FR=faculté
جامعة,N+Lieu+FLX=قارة+FR=université
مكتبة,N+Lieu+FLX=قارة+FR=bibliothèque

Table 1. Extracts of Arabic dictionaries

d. Predicate Representation of the Relations

Recognized explicit relations are represented using First Order Logic in the form of a Predicate (i. e. Relation Name) with two arguments: Person Name as the First Argument and the ORG Named Entity as the second Argument. The Name of the relation is the Lemma extracted from the Function Dictionary.

Example:
أمين عام (أبو الرب ، جمعية كليات الحاسبات والمعلومات)
Secretary-General (Abu al-Rub, Association of Colleges of Computing and Information)

For the implicit relation we generate systematically the name "Belong_to" ينتمي_إلى.

Example:
ينتمي_إلى (القاسم علي القاسم ، كلية الزراعة)
Belong_to (Al-Kacem Ali Al-Kacem, Faculty of Agriculture)

e. Translation of the relations

The translation of the recognized relations is done in the perspective of a multilingual assimilation of the analysed texts or documents. And the target language is the French but we can add other languages easily.

So we do not pay a great attention to the quality of the translation process, and we translate only lemmas of the words (i. e., nouns and adjectives) and not their derived forms as they occur in the text. For example:

جامعات Universities is translated as جامعة University,

عامة Générale is translated عام Général (masculine form).

French corresponding lemmas are added to the entries of the different Dictionaries that are used in the recognition process.

This translation, although it is not very good, allows us to question an Arabic text using French language and get answers in Arabic or even in French.

Example:

For the question: *Qui est le Président de l'Université Al-Albeit?* (see figure 4), we get the answer in Arabic: الأستاذ الدكتور الشوقفة and in French: Professeur Docteur Al-Chaouakfa.

5 Experimentation and Evaluation

For the system evaluation, we created a new corpus composed of the learning corpus and Wikipedia pages.

Figure 3 gives a sample execution of the proposed recognition system given by NooJ platform on the indicated corpus. It represents the recognised Relations between Person Names and ORG Named Entities with its corresponding Predicate representation.

Figure 3. A sample of the concordances obtained from the test corpus

Figure 4 gives the recognized relations with their French translation.

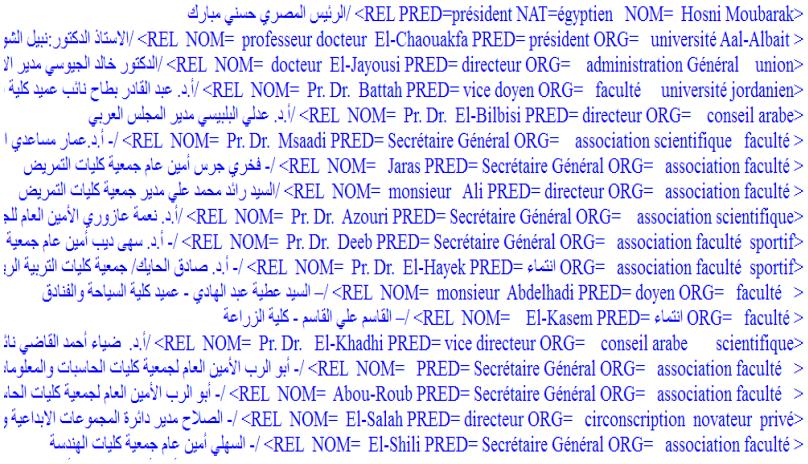


Figure 4. A sample of the concordances with their corresponding translation

The evaluation metrics we used for the recognition process, are Recall, Precision and F_measure (2*P*R/(P+R)). Let's remember that the recall measures the quantity of relevant responses of the system compared to the ideal number of responses; Precision is the number

of relevant responses of the system among all the responses he gave and the F-measure is a combination of Precision and Recall for penalizing the very large inequalities between these two measures. The values obtained in the evaluation of our work are:

	Precision	***Recall***	***F-measure***
Sports venues	63%	78%	70%

The values obtained can be explained by the fact that our dictionaries are not very large, especially for First Names and Last Names, and also by the fact that the names of organizations are very long andcomplex.

6 Conclusion

In this paper, we have presented a rule based approach for the extraction of functional relations between ENAMEX and ORG Arabic Named Entities. We particularly highlighted the NE relation Discovery problems. Some of them are specific to the Arabic language. These problems have been largely resolved, but some merit special consideration. The approach was implemented using NooJ platform. We have also given experimentation on a journalistic test corpus. The experimentation and the evaluation results are satisfactory.

As perspectives, we are working on expanding the dictionaries, especially for First Names and Family Names with corresponding translations. Also, we are interested in new types of relations for the economic domain in order to recognize events.Finally we project to integrate our system in a question answering system as a component for factoid questions.

Keywords: Relations between Named Entities, Extraction Process, NooJ transducers, Relation translation

References

Agichtein E. and Gravano L. 2000. Snow-ball: Extracting Relations from Large Plain-text Collections. *Proceedings of the Fifth ACM International Conference on Digital Libraries*.

Ben Hamadou A., Piton O., Fehri H. 2010. *Recognition and translation Arabic-French of Named Entities: case of the Sport places*, CoRR abs/1002.0481.

César P., Juan P., Isabel S., Paloma M. 2009. The *UC3M team at the Knowledge Base Population task*.

Shaalan Kh. and Raza H. 2009, "NERA: Named Entity Recognition for Arabic". *Journal Of The American Society For Information Science And Technology*, Vol. 60, N° 8, pp: 1652-1663, August 2009

Culotta A., Mccallum A. & Betz J. 2006. *Integrating probabilistic extraction models and data mining to discover relations and patterns in text*. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Morristown, NJ, USA. +.

Culotta A. and Sorensen J. 2004. Dependency Tree Kernel for Relation Extraction. *Proceeding of ACL-04*.

Jun Z., Zaiqing N., Xiaojiang L., Bo Z. and Ji-Rong W. 2009. StatSnowball: a Statistical Approach to Extracting Entity Relationships. *18th international World Wide Web conference* (WWW 2009).

Hasegawa T., Sekine S. and Grishman R. 2004. Discovering Relations among Named Entities from Large Corpora. *Proceeding of ACL-04*.

Kambhatla N. 2004. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. *Proceeding of ACL-04*, Poster paper.

Miller S., Fox H., Ramshaw L. and Weischedel R. 2000. A novel use of statistical parsing to extract information from text. *Proceedings of NAACL*-00.

Santos D., Mamede N., Baptista J., 2010. Extraction of Family Relations between Entities.

Silberstein M. 2005, "NooJ's dictionaries*". Actes de la conférence internationale LTC*, 2005, Poznan, Pologne.

Silberstein M. 2009, Syntactic parsing with NooJ. *Proceedings of NooJ 2009*, Finite State Language Engineering, Touzeur, Tunisia.

Stevenson M. 2004. An Unsupervised WordNet-based Algorithm for Relation Extraction. *Proceedings of the 4th LREC workshop "Beyond Named Entity: Semantic Labeling for NLP tasks*".

Zelenko D., Aone, C. and Richardella A. 2003. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*. 2003(2):1083-1106.

Recognition of negative paraphrases in Spanish

Angels Catena[(1)], Judith Sastre[(2)]

[(1)]*Autonomous University of Barcelona,*
[(2)]*Inbenta Professional Services S.L. / Autonomous University of Barcelona*

Abstract

The aim of this paper is to propose a preliminary approach in order to identify negative paraphrases in Spanish in a question answering system.
This research is the result of the collaboration between the researchers at fLexSem laboratory (Autonomous University of Barcelona) and the company INBENTA (http://www.inbenta.com/) that focuses on developing technology for corporate semantic search, based on a Semantic Search Engine.

1 Introduction

Contrary of what happens with traditional search engines, online customer care platforms allow users to express their queries in natural language. Thus, these tools must deal with one of the essential properties of natural languages: the possibility for a given meaning to be expressed by very different syntactic, lexical or morphological means. As a result, it is necessary to develop an automatic system able to recognize all the questions related to the same request regardless of the linguistic means used to express it. Linguistic knowledge used in semantic search engines is based mainlyon dictionaries or ontologies built for a specific domain.
If it is true that most of the user's searches can be partly recognized by an analysis basically focused on the lexicon (that is to say, identifying full lexical units and their synonyms), this is not the case for sentences related to such grammatical meanings as negation. Generally, automatic recognition of negative constructions is done without taking into account all the formal means that customers can use to express negation.
The *f*lexSem Laboratory at the Autonomous University of Barcelona has recently started a research collaboration with Inbenta. This company sells number of products related with the NLP field like e-mail management, virtual assistants, search engine optimization etc. All of these services are based on a semantic search engine (called ISSE).

In this paper, firstly we will focus on paraphrastic relation; then we will briefly describe negative paraphrases and finally we will present how these sentences were recognized by the system until now and we will *outline* some conclusions and perspective for future developments.

2 Linguistic paraphrases of queries

Natural languages have a common property which is to express the same meaning by very different syntactic, morphological, lexical (even prosodic) means.

Obviously, since it involves implementing online customer care platforms or a an automated agent, and since searches are not made by keywords but in natural language, the system needs to recognize different means that will be used to make different requests related to the same information.
Let us consider a prototypical user query asking for not receiving more advertisements in his cell phone or at home:

1. *No quiero recibir más mensajes comerciales*

   "I don't want to receive commercial messages anymore"[1]

Actually, users can vary the way to express an equivalent or near-equivalent meaning. It is for this reason that when we checked the logs, we found different paraphraseslike the following examples:

2. *Quiero dejar de recibir mensajes*
   "I want to stop receiving messages"
3. *No quiero que me manden más publicidad*
   "I don't want you to send me more advertising"
4. *Quiero que no me manden más mensajes comerciales*
   "I want you not to send me more commercial messages"
5. *Quiero que paréis de mandarme publicidad*
   "I want you to stop sending me advertising"
6. *No deseo publicidad*
   "I don't desire advertising"
7. *¿Cómo puedo no recibir más publicidad?*
   "What can I do to stop receiving advertising?"
8. *No me manden más mensajes comerciales*
   "Don't send me more commercial messages"

The goal of the ISSE is to perform matching between every possible user question and the content of the prototypical query showed above (example 1) regardless of the formal means employed.
On the other hand, examples (2-8) show that negation is a meaning that allows users different ways of expression so that it must be tackled from a multiple perspective. We will see in the next section that it triggers some complex lexical, syntactic and semantic relations that can cause problems for automatic recognition.
Inbenta has provided us with logs of different projects stored in a text file of 208 MB so that we could analyze the linguistic data. Since we had to deal with a huge amount of data, we used Nooj platform to build several simple syntactic grammars to take out only the target structures and we have imported them to an Excel file.
In order to illustrate the importance of the negative structures, we have taken out almost 198,000 occurrences of the word *no* that is the basic way to do negations in Spanish and about 26.000 occurrences of the word *sin* (without). It means more than 225,000 negative sentences altogether without considering lexical negation.

[1] For better readability, we decided to convey the meaning in English rather than to keep a litteral translation.

3 Different types of linguistic negation[2]

A part from the particle *no* to express simple negation in Spanish, other negative quantifiers or some negative polarity items (*nunca, nadie, ninguno...*), this grammatical meaning is often expressed in Spanish by other negative operators. In this section, we briefly describe some of them.

First of all, let us consider the preposition *sin* (*without*) that can be followed by a noun, as showed in (9), to express a privative meaning equivalent to 'que no tiene' ('that doesn't have') as in (9'):

9. *¿Un socio sin tarjeta Club paga la misma cuota?*
   "A member without the Club card does he pay the same fees"
9'. *(Un socio) que no tiene la tarjeta Club…*
   "A member who doesn't have the Club card…"

The same preposition followed by an infinitive has a passive meaning:

10. *¿Dónde tiro las pilas sin usar?*
   "Where to throw unused batteries?"
10' *(¿Dónde tiro las pilas) que no se han usado?*
   "Where to throw batteries that haven't been used"

As showed in the example above, the phrase *sin + infinitive* is in a paraphrastic relation with the passive reflexive (10') but it could also be equivalent to the analytical passive (*ser + past participle)* with a preceding pronoun *que* (*¿dónde tiro las pilas que no han sido usadas?*).
When *sin* is followed by *que* it expresses logical relations related with conditional values as in the following examples:

11. *¿Se puede hacer una transferencia del vehículo sin que esté el propietario delante?*
   "Can we make a car transfer without requiring the car owner to be present?"
11' *¿(Se puede hacer una transferencia del vehículo) si el propietario no está delante?*
   "… If the car owner is no longer present"
12. *¿Dónde puedo vender un móvil sin que me cobren por ello?*
   "Where do I sell a cell phone without paying for it?"
12' *(¿Dónde puedo vender un móvil)pero que no cobren por ello?*
   "…but without them to charge me for this"

Some prefixes can also carry negative meanings[3]. The most frequent negative prefixes to express lexical negation in Spanish are *in-, des-, anti-, extra-* and the particle *no*followed by a noun:

[2] For a study of negation in Spanish see Sánchez López (1999).
[3] In this paper, the term "negative meanings" is used to refer to several grammatical meanings as opposition, reverse action or privative values.

13. *Reclamaciones por disconformidad con la reparación*
    "Complaints which involvea disagreement with the repair"
14. *Una solución antirrobo para la moto*
    "Anti-theft solution for the motorbike"
15. *Actividad extraordinaria*
    "Extraordinary activity"
16. *Cantidades impagadas*
    "Unpaid quantities"
17. *El incumplimiento de los requisitos*
    "Unfulfilment of requirements"
18. *Seguros para conductores no habituales*
    "Insurances for non-regular drivers"

In all these cases, the same meaning is carried by syntactic negation :

13' (Reclamaciones) porno estar conforme (con la reparación)
"… of not being satisfied …"
14' (Una solución) para que no me roben la moto
"…for not being stolen"
15' ... que no es ordinaria
"which is not ordinary"
16' (Cantidades) que no se han pagado; sin pagar…
"That haven't been payed"
17' No cumplir los requisitos
"Do not fulfil the requierements"
18' (Seguros para conductores) que no conducen habitualmente
"…don't drive regularly"

Sometimes, negation is included in the meaning of some lexical units. The following sentences illustrate the paraphrases between some verbs and a presuppositional negation:

19. *¿Si cambio de seguro pierdo la bonificación?*
    "If I change the insurance will I lose the discount"
19 *(Si cambio de seguro) ya no tengo bonificación / Me quedo sin bonificación…*
    "If I change the insurance, will I not have the discount anymore"
20. *Dejar de recibir publicidad*
    "Stop receiving advertisements"
20'...no recibir (más) publicidad
"Not to receive advertisements anymore"

Some adjectives are also used with this meaning:

21. *Derechos de un antiguo cliente*
    "Rights of a former customer"
21' (Derechos de alguien que)(ya) no es cliente
"Rights of someone that is not a customer (anymore)"

However, in certain cases, adjectives that include negative meanings combine with deontic modality:

22. *¿Los jubilados están exentos de pagar el impuesto inmobiliario?*
    "Are pensioners exempt from paying the property tax?"
22' *¿(Los jubilados) no tienen que pagar?*
    "Pensioners don't have to pay"
23. *¿Quiénes están eximidos de ayunar?*
    "Who is exempt from fasting?"
23' ¿Quién no tiene que ayunar?
    "Who need not to fast?"

Finally, other adjectives like *zero* can have a meaning equivalent to the indefinite adjectif *ningun*oand *pendiente de*followed by a noun or an infinitive can function as an aspectual negative polarity termas well:

24. *¿Tengo descuento si he tenido cero accidentes?*
    "Do I have a discount if I have had zero accidents"
25' *(¿Tengo descuento) si no he tenido (ningún) accidente(s)…*
    "Do I have a discount if I have had no accident"
25. *¿Dónde puedo ver los documentos pendientes de aprobar?*
    "Where can I see documents pending to be accepted"
25' *(¿Dónde puedo ver los documentos) sin aprobar / que no (han sido) aprobados (todavía)?*
    "Where can I see documents that have not yet been accepted?

The meaning of some verbs can be paraphrased as 'to cause that something ends' which entails a terminative meaning :

26. *Quiero cancelar (rescindir, suspender, anular, darme de baja, etc.) del seguro*
    "I want to cancel (rescind, hold on, etc.) the insurance"
24' *… no quiero renovar el seguro*
    "I don't want to renew the insurance"

We found in the corpus many other ways to express negation sometimes combined with other grammatical meanings. In the following example, translation allows toraise negative meanings conveyed by several lexical units and that wouldappear innear-equivalent Spanish sentences:

27. *coche ajeno* ("car that is not mine"); *librarse de pagar* ("be freed from paying"), *estoy fuera del país* ("I'm not in the country"); *los papeles que me faltan* ("I haven't got the papers"), *una dirección distinta de la del DNI* ("an address that is different from the one ID")…

In addition to the previous occurrences, we must take into account the possibility for certain negated lexical units that are not inherently negative, to establish paraphrastic relation with their antonymic counterpart :

28. *He olvidado la contraseña*
    "I forgot the password"
28' *No recuerdo la contraseña*
    "I don't remember the password"

4 The Semantic Search Engine

In this section, we will briefly present the Inbenta Semantic Search Engine (ISSE) performances and the way it operates in order to resolve – or try to resolve – up to now the problem of paraphrases.

As we have mentioned above, the aim of the system is to perform matching between two questions: the one posed by the user and the prototypical question manually created by the editor. Three modules have been developed with this end: logs, lexicon, contents and performance test.
The lexicon module contains three lexical databases that store on the one hand, the general lexicon (e.g. *vehicle*), on the other, the lexicon related to a specific domain (e.g. *redeem the mortgage*), and finally, terms created for a particular project (e.g. *punts estrella* / "starpoints").
In turn, every lemma (as shown in Fig.1) contains lexicographic information mainly for part of speech, synonyms and syntactic derivations (nominalizations and verbalizations)[4]. Every part of speech has a priori a semantic weight. For instance, in a sentence as *to make a credit transfer* the noun has a high value but the verb *to make* has low weight. Some other lexical units like the verb *desear* ("to like"), numerals, adverbs, etc. are automatically removed from the sentence.

Figure 1. First module: lexicon

[4] Lexical relations are labeled using the lexical functions **Syn, S0, V0**… For more information we refer the interested reader to Mel'čuk (1996, 1998)

The second module (Fig.2) comprises a set of predetermined answers related to some prototypical questions. When a user types a query, the system should suggest one or several answers chosen among these pre-established contents.

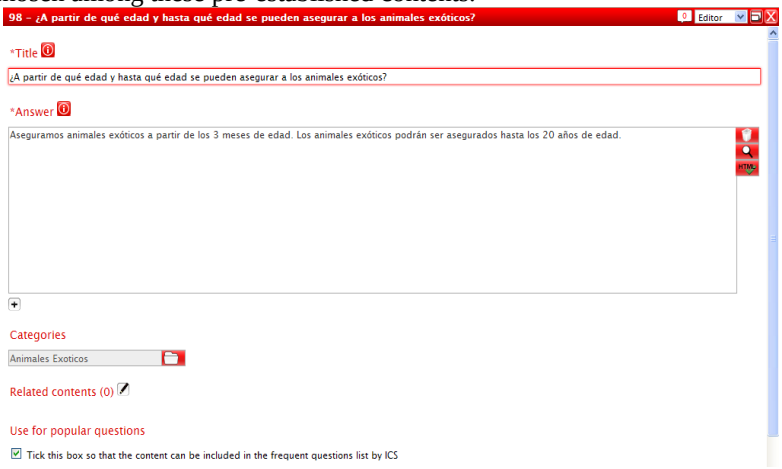


Figure 2. Second module: contents

Finally, the test module allows linguists to see how the matching is done. For instance, they can see which was the semantic weight attributed to every recognized lexical unit (fig. 3). It shows as well relations that were applied (synonyms or other lexical functions, etc.). We must emphasize that up to know the system is strongly lexically-based and it gives one or more answers that contain the lexical units (or their synonyms) found in the sentence. Of course, it is a good solution when there are few contents related to one or two keywords but if it is not the case, usersmight have several non relevant answers to their queries or none at all.

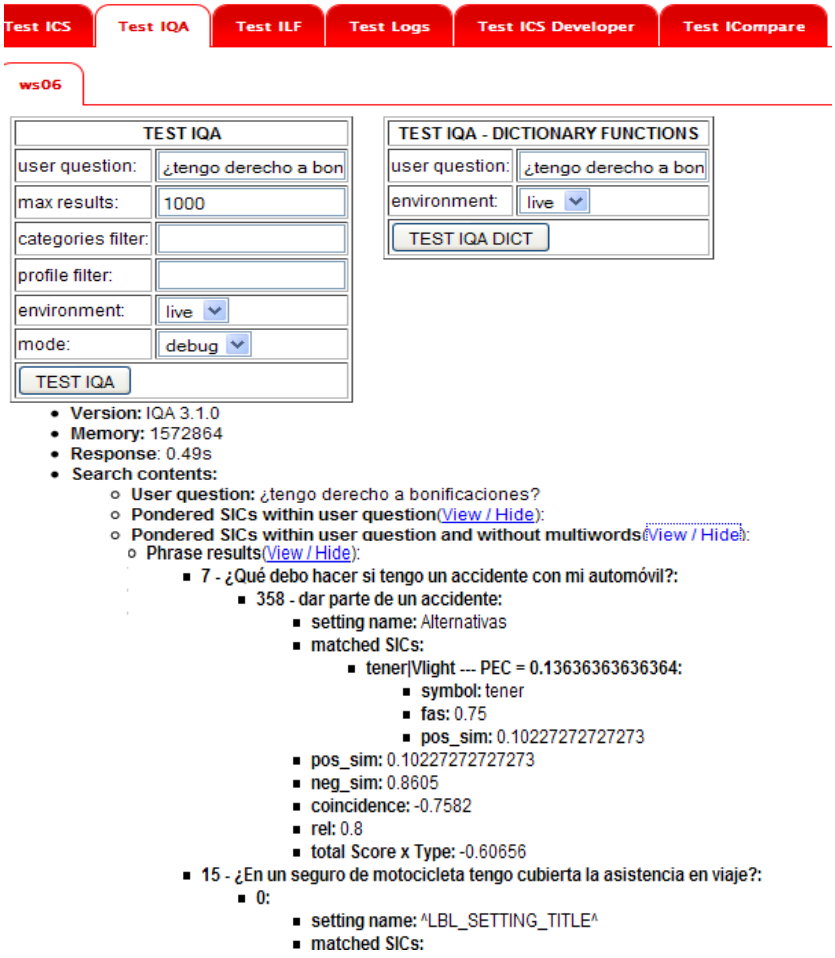


Figure 3. Third module: performance test

As far as the weigthing coefficient is concerned, the word *no* has always the same value in the lexical database. However, the word usage examples belowshow that although its meaning is empty in some interrogative sentences or combined with some modal verbs (29), in many other cases, its deletion might change the meaning of the sentence (30):

29. *(No sé) qué es seguro a terceros*
    "I don't know what a third-party insurance is"
    (No puedo) acceder a mi cuenta
    "I can't get access to my account"
    ¿(No) es posible votar desde el extranjero?
    "Is it not possible to vote from abroad"

30. *Devolver un importe cargado que no es un recibo*
    "Return a charged amount that is not an invoice"
    No me han cobrado la renovación

"They didn't charge me the renewal"
No recibir sms en el móvil
"No receiving sms on cell phone"

5 Conclusions and future work

So far, we have presented different ways of expressing negation in Spanish and we have established paraphrastic relations between all these means. The goal of such a description is to improve the negative paraphrases recognition by the ISSE.
It is clear that after the first approximation, finer results will be needed for some applications.
At this point, we can state that some antonyms can be easily related in the lexical database through the lexical function **Anti**[5].However, this kind of equivalence only with contradictory antonyms (as in examples 28 and 28'). Formalization of contrary antonyms (e.g. *aconsejar* /"to advise" and *desaconsejar* / "to advise against") should take into accountthe ambiguity of the negative counterpart (e.g. *no aconsejar* can be understood as *not to advise doing something* or *advise against something*) or understatement for antonyms that carry a "more or less" meaning as scalar adjectives (e.g. *no es barato* / "it is not cheap" may mean *es caro* "it is expensive").
The next step of our work is to solve some problems that remain uncovered concerning the scope of the negation and syntactic ambiguity (in the sentence *¿Cómo puedo obtener un listado de talleres sin teléfono?* / "How can I get a list of garages without telephone?" the prepositional phrase *sin teléfono* can be a complement of the main verb or a complement of the noun *talleres*)or the ellipsis of the verb in some coordinated clauses (e.g. *Me funcionan las tarjetas en compras perono en el cajero* / "My credit cards works on shops but not in ATM").
Likewise, we must be careful with frozen expressions that cannot be paraphrased by the negative counterpart in a natural way (e.g. *¿Puede hacerse un sistema de alarmas sin cables?* / "Can we make a wireless alarm system?").

References

Mel'čuk, I. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: Benjamins, 37-102.

Mel'čuk I. 1998. Collocations and Lexical Functions. In A.P. Cowie (ed.): *Phraseology. Theory, Analysis, and Applications*, Oxford: Clarendon Press, 23-53.

Milićević, J. 2007. *La paraphrase. Modélisation de la paraphrase langagière*. Bern: Peter Lang.

Sánchez López, C. 1999. La negación. In I. Bosque y V. Demonte. (dirs.), *Gramática descriptiva de la lengua española*, Madrid: Espasa. Calpe, 2561-2634.

[5] For a description of paraphrasing rules using the FL Anti, see Milićević (2007).

Colour semantics and ambiguity, processing approaches with NooJ

Marcel Puig Portella
Autonomous University of Barcelona, Spain.
Marcel.Puig@uab.cat

1 Introduction

Computational Linguistics is a complex multidisciplinary study field constantly challenged in its practical applications by the growing industry of digital text translation and processing. One of the greatest difficulties facing linguistic processing is facing is the question of textual ambiguity.

It is not an obvious or solved question. Anything related to Natural Language analysis and process handles with complex combinatory semantics processing. Different engines provide a diversity of approaches, but independently of their perspective, soon restrictions and limits appear, narrowly related with the interacting capacity of the program with the different text forms.

In this article our treatment of linguistic structures follows a very specific field and pathway, focusing in those particularly ambiguous wordforms in relationship with the semantic of colours. For linguistic software, to detect words like *orange, black, brown*, etc., means to be facing an ambiguity misguiding its semantics and the general text comprehension: Almost any colour-related wordform can refer at the same time to other semantic fields, having a completely different syntactic behaviour.

Our work finds its roots in the *Sens-Texte* theory, using at the same time NooJ as linguistic processing tool.

2 Analysis

Background

The *Sens-texte* theory refers to the Igor A. Mel'čuk's linguistic model in processing the different meaning layers of a text. A diversity of applications has already been demonstrated and continuously emanate from this theory. In Mel'čuk's own words, his modelling theory rather than a construct was originally created in order to lighten the rather complex world of word selection and combinatory semantics of any natural language. The transferring process of a semantic meaning into a combinatory text sequence is what the theory's name arises from.

Mel'čuk's theory takes into account different deeps of analysis interacting among them to recreate and transfer a semantic meaning.

NooJ is the Max Silberztein's linguistic-processing software program, finding its roots in the previous work of INTEX. It permits very complex searches in a multiple levels text consideration and analysis.

NooJ is a free software available in internet under the address www.NooJ4nlp.net/. Modules for different languages are already available, while others, for more languages are in progress.

In our work, we are also indebted with the previous work of Gaston Gross about the *classes d'objets*, whose main lines show the many specific subdivisions a simple wordform can be subjected after its simplest syntactico-semantic combinatory properties.

3 Ambiguity in linguistic computer processing

In natural language (NL) translation and processing, ambiguity is one of the highest problems a software handles with. Ambiguity refers mainly to any homographic situation, in its largest sense, vehiculating a different meaning or a diverse syntactic combinatory. Already various ways have been proposed to solve that necessary problem:

Cluster-based, probabilistic, translation has its effective support in statistic correspondences among wordforms, connected by a complex network of cooccurrences or meanings.

Corpus-based translation, on the contrary, operates through another connexion network, a one created throughout a linguistic modelling whose relationships are not fixed following mere textual coocurrences but considering individual of any implied particle or wordform, in order to lately form structural regularities with prediction capacity.

One of the implicit characteristics of corpus-based NL processing is the previous need of corpus dictionary.

3.1 Our case: colour homograph processing

Our work with NooJ is based on colour potentially related wordforms in Catalan language, lately to be used in translation and extended to other languages.

An explanation for the special ambiguity of those specific forms could probably be traced into their etymology, a concept that computational processing will difficultly handle with.

Even probabilistic values, like plausibility, change radically depending on almost any context for such wordforms. Speaking in terms of etymology, in most colour-related wordforms, the semantic of the colour appears in second place, even if their plausibility could have changed throughout the times: many common, more or less concrete, perused objects are mostly at the origin of the term. Fruits (*orange*, *lemon*), as for instance, domestic plants or minerals (*emerald*, *ruby*), directly or throughout borrowing from other languages (*cyan*, *crimson*), were the first meanings those words were referring.

This specific characteristic, together with the shifting of their plausibility translates into an especially difficult computer processing. The question could be shown throughout examples, the main point being mainly how to "teach" a software program to differentiate between text sequences like *the orange shirt* and *the orange squeezer*, just considering that, in other languages, the *term* orange could have a completely different translation, as well as very different combinatory properties.

Articles have already been written about the implications of any text tagging process, some of them trying to implement their semantic coverage to previously processed texts.

Any intention of creating a really working semantic network should at least take into account plainly consistent facts:

-NL processing will always be facing new, and therefore untagged, texts.

-Text reprocessing, i.e.: tagged texts processing, can prove useful in software testing. Attempts shouldn't be brought so further as to properly believe that tagged texts processing

is *per se* a real disambiguation process. Automatic text processing as well as translation should recognize the challenging reality of untagged, fresh, texts and corpora as its principal defy.

3.2 Nooj color-disambiguation work

In our little work on colour-semantically related forms, and as already said, we propose to solve the ambiguity problem throughout semantic networks of restrictive combinations.

In NooJ this is known as a Grammar. NooJ Grammars find their basis on a previous corpus processing and tagging of lemmas for the creation of a dictionary of lexical entries. It is on those entries, as categories, that the selecting network will be able to be built.

A Syntactic Grammar is an artificially created network of relationships between different categories and simple wordforms that only requires two requisites:

a) The previous tagging of different wordforms for the building up of a dictionary of entries.

b) The creation of new, contextually restricted, relationships between them.

The grammar can be seen and operated under a tree form of directional vectors processing the individual meaning of any single wordform into a more complex, general, meaning, as well as restricting their own collocation. Sentences in Catalan like *plou a bots i a barrals* would be able to be translated into a less literal form, like *it rains cats and dogs*, or into a less marked form like *it rains a lot*. Thus, single wordforms, when necessary, are processed together into more complex semantic units.

On processing colour-related wordforms with NooJ, we met the question of the productivity of a specific NooJ grammar for their disambiguation. In other words: are syntactic fixations in colour wordforms regular and productive enough as to consider the creation of disambiguating grammars specifically for them? This question will be answered in the following solutions we propose for the ambiguity problem while creating new grammars:

3.3 Colours and gender disambiguation:

A gender selecting grammar is much easier to create than a syntactic grammar, with the advantage of its potential productivity. In Catalan, like in most roman languages, gender disambiguation is very productive and persistent throughout different number variations, as well as easily visible throughout the article, adjectives and other determiners.

The colours are always masculine in Catalan, speaking of terms with a nominal syntactic behaviour. The problem of such grammars would be related to a productivity linked to a singularity-dependence: disambiguation is only possible when a given semantic possibility has a different gender from its homograph, as in the case *el taronja* (colour-related) vs. *la taronja* (fruit), never in examples like *el rubí* (colour-related?) vs. *el rubí* (mineral?).

3.4 Colours and number disambiguation:

The number is a mostly universal linguistic pattern, with a great potential in disambiguation productivity. However, disambiguation has two different and necessary edges: detection and use.

In that specific case, detection would be helped by the extremely regular number inflection in Catalan (mostly with the single addition of an –s to the initial masculine or feminine forms).

Nevertheless number grammars are not, even in specific cases, as neatly strict in sequential disambiguation as gender grammars are. Explained in a practical way: colour terms do not so frequently appear in their plural forms as their homographs do, but they sometimes do.

Therefore, we will here be mixing two different considerations in the linguistic processing: one corpus-based and the other working with concepts like plausibility.

The remaining of corpus directly unprocessed words plus the non negligible number of possible errors made us discard this option as a suitable disambiguation grammar.

3.5 Colours and the lexical entry:

Even before the advent of computational linguistics, the use of a new lexical entry was a better form of disambiguation for homonyms.

For frozen sentences, those fixed independently of their context, the creation of a new lexical entry can be the best solution. A theory, even a linguistic model, remains apart of such a "practical way" consideration; even if many lexicological theories already agree in the consideration of homonyms (words with the same form but with different meaning and combinatory syntax) as different lemmas requiring separated lexical entries.

In practical work, the lexical entry becomes the easiest form of disambiguation for computer processing. Complex wordforms like *green house effect* or *rose-colored glasses* are easily and sequentially recognizable in any context. For such cases, semantic tagging comes automatically with the recognition, with no need of algorithms or further syntactic consideration.

Moreover, the forms this way recognised are universally stable in almost any context. The clear advantage of this method is its certitude in disambiguation and its absolute productivity, even restricted in the numbers of relative occurrences and terms. The inconvenient is the real impossibility of its extension for the disambiguation of other, non fitting, wordforms.

Notwithstanding its restrictions, it is probably the solution for compound disambiguation, in terms like: *black bear, white eagle, red skin*(in Catalan *pell roja*, not **pell vermella*), etc., as well as for frozen sentences, independently of their length or conceptual description.

Owing to its limitations, it is also worthless in the treatment of non-affixal derivation, in those words whose meaning is narrowly determined by other words or determiners in sequences that are not properly frozen, like in *bright orange*.

3.6 Colours and syntax

It is the field the name of Syntactic Grammar arises from. And, in one sense or another sense, the term combinatory becomes the keyword in any case of corpus-based processing.

As already stated by the *Text-meaning* theory, different combinations operate into different meanings, in production as well as perception. In general, speaking of computational linguistics applied to NL text corpora, the only semantics susceptible of

processing will be the graph-explicit one. In our opinion, semantics will be able to express throughout three main graph characteristics:

Sequentiation
Difference
Combination

Text *sequentiation* is already known from the older, linear definition of the word language. In our case, it refers to the way language is always computationally processed, at least in its reading as well and in the rear end –in reprocessing semantics into new meaningful sentences- independently of the possible middle stages.

Of text or graph *difference* we have already given examples when analysing colour disambiguation throughout gender and number grammars.

And *combination* will have henceforth our attention in disambiguation, specifically referring to its sequential aspects, syntagmatic and non-paradigmatic.

While real *frozen forms* present themselves under such restrictive combinations that no further operation or modification is allowed in transferring their meaning, other, less restrictive, sequential combinations are more free, allowing nevertheless to track and restrict the meaning or their single constituents.

NooJ software can help us process those syntactic meanings. Not without restrictions:

Difficulties in processing syntactico-semantically restricted sequences are mostly related to a natural quality of NL text processing: *variation*. Variation can be the result of many other characteristics: graph qualities (orthography, as for instance), dialectal differences as well as any other context-independent, and therefore unpredictable, change capable to disturb regularity or its extension into the text computational processing.

Syntactic grammars in text disambiguation

The following examples of syntactic grammars in Catalan, mainly applied to verbal forms, are a first intent of formalisation for natural language sequences containing ambiguous colour semantically-related wordforms:

Previous step, the macrocategory <Colour>

The creation of the macrocategory <Colour> (with its first graph in capitals) will allow us to process all the susceptible colour wordforms we have previously put in a dictionary, with all their respective inflections, simultaneously. This application would have no sense in a production way, as for instance in order to evolve active inflected forms; however, we are just looking for the creation of a macrocategory able to recognise very restricted wordforms susceptible to behave as colours.

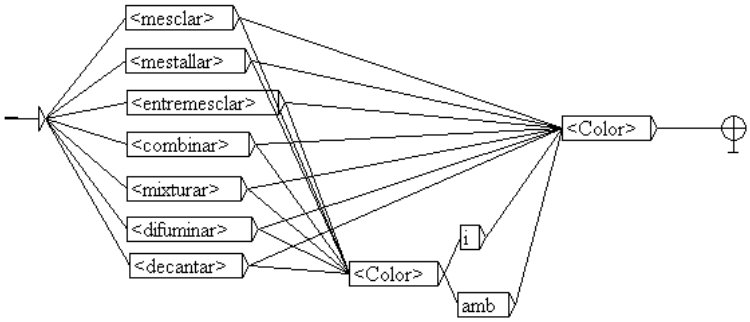
The next step is the creation of the proper Syntactic Grammar, grammars with colour-selecting verbs:

Verbs are probably the greatest potential disambiguation tool for their selected object:

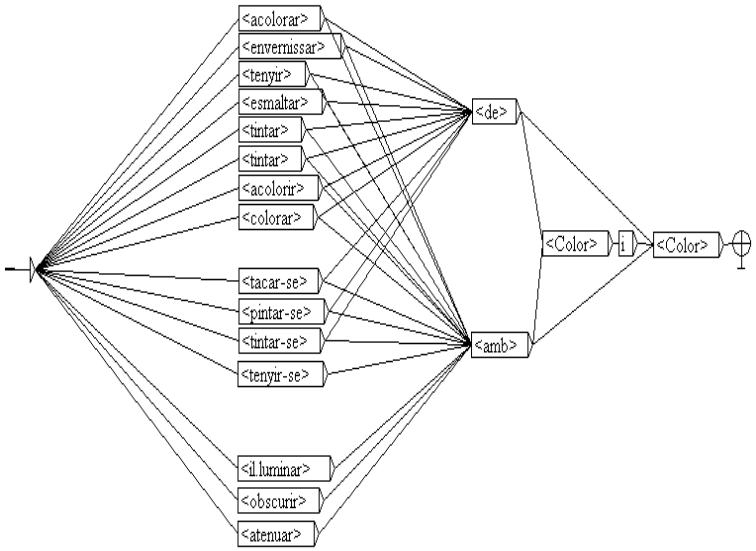
Objects can be directly (with no other wordforms in between) linked to the verbal form, but are most frequently seen as introduced by a preposition, like in the examples:

Il.luminar (amb) Eng=to lighten (with/
*Acolorir (de / amb) Eng=to colour (with/ *"of")*

NooJ syntactic grammars offer the possibility of considering any wordform as a potentially selected candidate, optionally allowing its discard and further processing (as it sometimes becomes convenient with the prepositions), so as to allow the simultaneous process of multiple options.

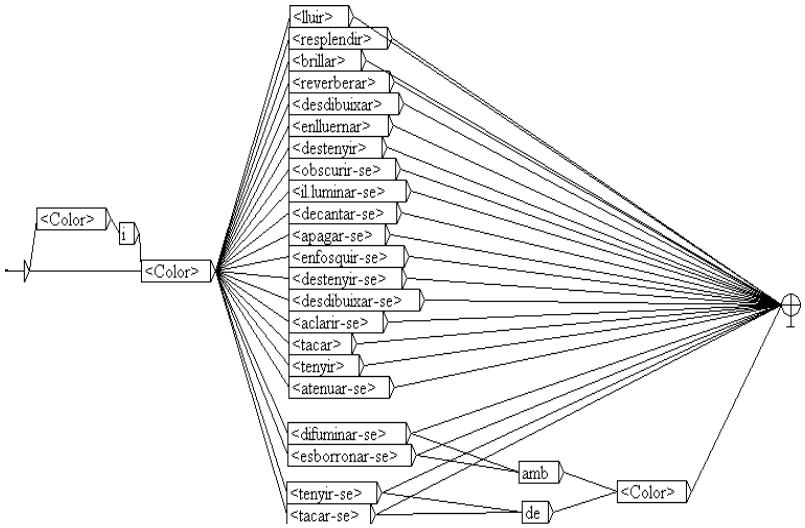


The same operation gives us the possibility of simultaneous disambiguation in sequences with more than one colour wordform:

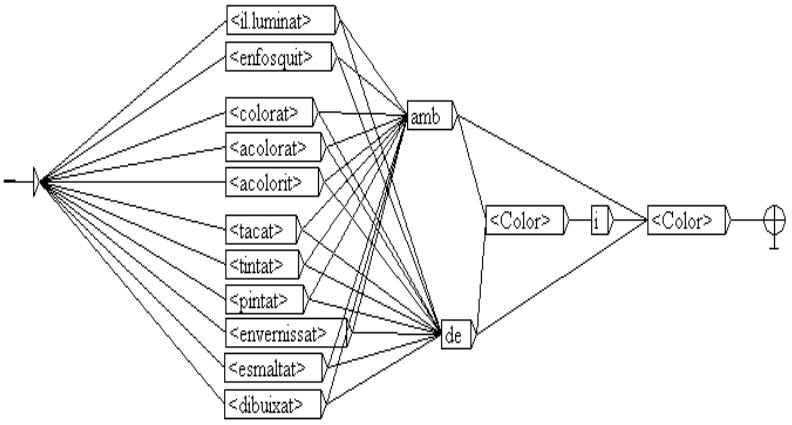


The examples show the normal presentation of the exposed sentences. In specific cases, particular consideration must be drawn the same way to the optional presence of articles as well as other possible determiners.

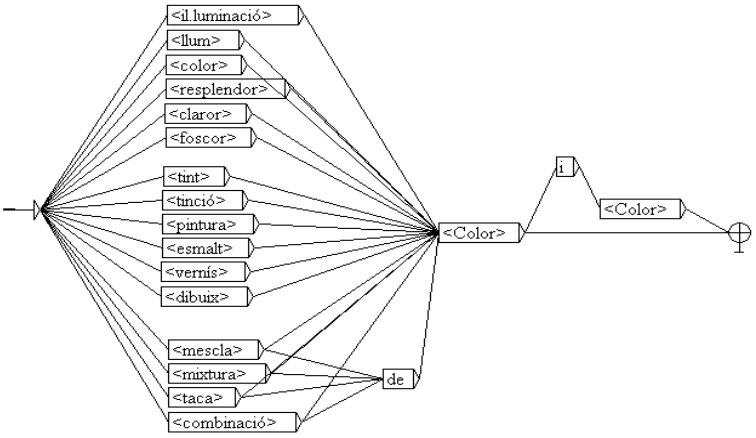
Colour semantically-related wordforms can also behave as subjects:



Furthermore, verbal forms can be seen as participles, or adjectivalised. In such cases, the object, even being computationally processed the same way, would traditionally be censed an adjectival complement:



Finally, here is the example of some nominal forms having a verbal origin and accepting adjectival complements:



Colour disambiguation is, in those procedures, a double edge tool: Its disambiguation capacity always depends on the semantics left in the "other" non selected option, the "other split half". This way, these local grammars can be seen just as collector selecting tools, with two possible options:

a) The given wordform fits into them, in which case it is selected for a particular meaning.

b) The wordform doesn't fit in the context selecting local grammar, in which case, it usually and automatically becomes a candidate for its former, single, usually context-independent meaning; or, for the case of cluster linguistic processing, for the most plausible next one.

4 Conclusions

In this article we intended to present a general overview of the problematic of ambiguity in natural language text processing via some examples in the especially difficult field of colour-related wordforms in Catalan language.

We started with the description of the problem, the background and some previous attempts of approach as well as of our theoretical basis and linguistic processing tool, NooJ that allowed us the simultaneous process of entire masses of corpora, for creating our dictionary and generating our Syntactic Grammars.

Although, for different NooJ modules, numerous syntactic grammars have already been developed, they usually have very specific applications, format-like recognizing or for specific data input. The use of Syntactic Grammars as a tool in linguistic disambiguation facing unformatted texts is a potential large field still being developed.

In our work we just intended to expose some possible uses of the tool in the disambiguation of the especially ambiguous wordforms related to the semantics of the colours in Catalan. We propose to use Syntactic Grammars in the way of a selecting network but with the additional insertion of new macrocategories, <Colour> in our case, in order to specifically disambiguate the lexically selected words.

Further work would be necessary to formalize entire masses of texts and to extend such applications to regular networks capable to break general homograph ambiguity in the computational processing of text corpora.

Key words: Colours, NooJ, linguistic processing, Catalan language, homograph disambiguation.

References

Blanco, Xavier; Silberztein, Max. Eds. 2008. *Proceedings of the 2007 International NooJ Conference* (Barcelona). Cambridge Scholars Publishing

Gross. G. 1992. Forme d'un dictionnaire électronique.*Actes du colloque La station* de *traduction de l'an 2000.*

Gross G. 1994. Classes d'objets et description des verbes. *Langages 115*. Larousse.

Mel'čuk, I. & Polguère, A. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Louvain-la-Neuve, De Boeck.

Mel'čuk, I. 1998. *The Meaning-Text Approach to the Study of Natural Language and Linguistic Functional Models* (Invited lecture) In S. Embleton (ed.): LACUS Forum 24, Chapel Hill: LACUS, 3-20.

Silberztein, M. *2003. NooJ manual.*

Silberztein, M.1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX,* Paris, Masson.

Automatic Transformational Analysis and Generation

Max Silberztein
LASELDI, Université de Franche-Comté
max.silberztein@univ-fcomte.fr

Abstract

I present a new automatic transformational engine for NooJ, capable of producing all the paraphrases of any given sentence (elementary or complex). The new engine does not require the implementation of a new level of linguistic description, as it uses slightly enhanced traditional NooJ syntactic grammars.

Introduction

NooJ is a linguistic development environment: it allows linguists to formalize several levels of linguistic phenomena: orthography and spelling, lexicons for simple words, multiword units and frozen expressions, inflectional and derivational morphology, local, structural and transformational syntax. For each of these levels, NooJ provides linguists with one or more formal tools specifically designed to facilitate the description of each phenomenon, as well as parsing tools designed to be as computationally efficient as possible.[1]
At the transformational syntactic level of analysis[2], NooJ allows linguists to associate a given sentence with a paraphrase via an elementary transformation, e.g.:
[Pron-0] *John ate an apple = He ate an apple*
[Passive] *John ate an apple = An apple was eaten by John*
as well as to associate more than one sentence with a complex paraphrase via a complex transformation, e.g.:
[Coord-1] *John ate an apple; John ate a pear = John ate an apple and a pear*
In NooJ, these transformations can be programmed with grammars such as the following one:

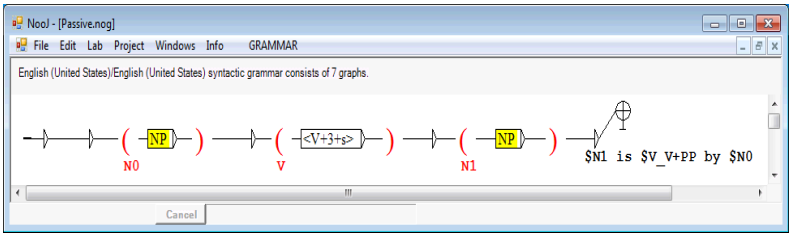


Figure 1. The [Passive] elementary transformation

[1] This approach is the opposite of that of most computational linguistic tools such as XFST, LFG, HPSG, etc. which provide a single formalism supposed to be used to describe everything.
[2] I am using the term *transformation* as in (Harris 1968) and (Gross 1975) : an operator that links semantically equivalent paraphrases, as opposed to (Chomsky 1957) whose transformations link deep and surface structures.

In this graph, the transformation's arguments are stored in variables ($NO, $V and $N1); for instance, when parsing the sentence *John loves Eva*, the variable $N0 would store the linguistic unit *John*, $V would store *loves* and $N1 would store *Eva*. The grammar's output "$N1 is $V_V+PP by $N0" would then produce the string *Eva is loved by John*. Morphological operations, such as "$V_V+PP", operate on linguistic units rather than plain strings: NooJ knows that the linguistic unit *loves* is a form of the verb *to love*, and it has access to all the linguistic information associated with this linguistic unit: conjugation paradigm, derived forms, syntactic and semantic properties. NooJ is therefore capable of processing the variable $V and applying to it the operator "_V+PP" which stands for: inflect the verb in its past participle form.

This system has been used quite successfully in bilingual Machine-Translation systems, see for instance (Ben Hamadou et alii, 2010) and (Fehri et alii, 2010) for Arabic-French translation, (Barreiro 2008) for Portuguese-English translation and (Wu 2010) for French-Chinese translation.

Any serious attempt at describing a significant part of a natural language will involve the creation of a large number of elementary transformations:

[Pron-0]	*John eats an apple = He eats an apple*
[Pron-1]	*John eats an apple = John eats it*
[Pron-2]	*John gives an apple to Marie = John gives her an apple*
[Progr]	*John eats an apple = John is eating an apple*
[Preterit]	*John eats an apple = John ate an apple*
[Impfct]	*John eats an apple = John has eaten an apple*
[Futur]	*John eats an apple = John will eat an apple*
[Cond]	*John eats an apple = John should eat an apple*
[Passive]	*John eats an apple = An apple is eaten by John*
[Negation]	*John eats an apple = John does not eat an apple*
[Cleft-0]	*John eats an apple = It is John who eats an apple*
[Cleft-1]	*John eats an apple = It is an apple that John eats*
[Question-0]	*John eats an apple = Who eats an apple?*
[Question-1]	*John eats an apple = What does John eat?*
[Question-V]	*John eats an apple = What does John do?*
[Nom-0]	*John loves apples = John is an apple lover*
[Nom-V]	*John loves apples = John's love for apples*
[Nom-1]	*John gave the card to Mary = The card is John's gift to Mary*
…	

Figure 2. Basic transformations

In NooJ v2.x, each of these transformations would have to be described by one graph… actually *two* graphs, because each pair of paraphrases involve two different (reverse) constructions: one wants to be able to produce not only *An apple is eaten by John* from the sentence *John eats an apple*, but also the sentence *John eats an apple* from the sentence *An apple is eaten by John*.

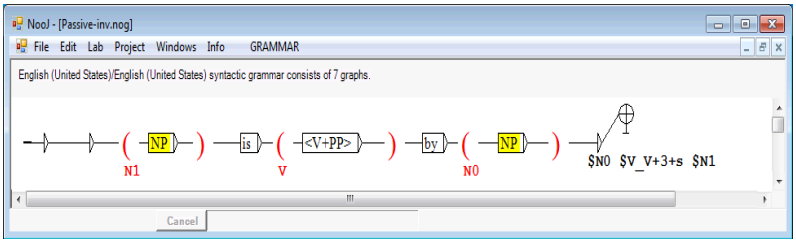


Figure 3. The [Passive-inv] elementary transformation

Unfortunately, even two graphs per transformation would not be enough to produce all the paraphrases of a given sentence, because most of the transformations presented above can be combined with other ones. For instance, from the sentence *John eats an apple*, one can construct the question *What did John not eat?* by combining three transformations: [Question-1], [Preterit] and [Neg]. These combinations of transformations cause the number of pairs of potential paraphrases to explode: for 100 elementary transformations, one would need to construct up to 10,000 pairs of graphs to represent all possible combinations of two transformations, one million pairs of graphs for three transformations, one hundred million pairs of graphs for four transformations, etc.
Clearly, this solution is not adequate.

Chains of transformations

Rather than describing combinations of transformations, linguists have traditionally worked on chains of transformations. For instance, here are two chains of transformations:
John sees a dog **[Cleft-0]**→ *It is John who sees a dog* **[Neg]**→ *It is not John who sees a dog*
John sees a dog **[Neg]**→ *John does not see a dog* **[Cleft-0]**→ *It is John who does not see a dog*
Notice that the order in which transformations are applied is important. As a matter of fact, ordered chains of transformations are often presented as *the* result of a transformational analysis of a sentence. In other words, the very meaning of the sentence *It is John who does not see a dog* is represented by the following formula:

*John sees a dog***[Neg] [Cleft-0]**

In principle, chains of transformations would allow an automatic system to produce any potential paraphrase from a given sentence with a large economy of scale because the system would only store elementary transformation graphs: no need to construct a separate graph for each of the numerous possible combinations of transformations.
Unfortunately, this traditional approach creates at least three problems:

1 Complex elementary transformations

It is not possible to design an elementary transformation graph that could be applied to a sentence independently from the sentence's structure. For instance, even such a simple transformation as **[Neg]** is not applied the same way on an elementary sentence and on a complex (i.e. transformed) sentence:
John sees a dog **[Neg]** → *John* does not *see a dog*

John does shopping **[Neg]** → *John* does not *doshopping*
Why does John see a dog ? **[Neg]** → *Why* does not **do John see a dog ?*
John is buying a car **[Neg]** → *John is* not *buying a car*
John is rich **[Neg]** → *John is* not *rich*
John has bought a car **[Neg]** → *John has* not *boughta car*
John has a car **[Neg]** → *John* does not *have a car* | *John has* not *a car*

To produce the correct negation for the sentence *John is buying a car* for instance, one must insert only the adverb *not* but not the verbal form *does*. To produce the correct negation for the question *Why does not John see a dog ?*, one must delete the extra occurrence of the verb *to do*. Sentences with *have* (as well as *need* and *must*) accept two negative paraphrases. Sentences in the futur, in the conditional or in preterit would each require different transformation graphs, etc.[3]

In conclusion: in order to construct a system capable of performing elementary transformations in cascade, one would need to design transformation graphs that could produce paraphrases correctly, not only for elementary sentences, but also for complex (i.e. already transformed) sentences. Such graphs would necessarily include the structure of all their possible inputs, *i.e.* the initial elementary sentence as well as all its possible transformed paraphrases. In effect, building such graphs would not be significantly less complex than building a pair of graphs for each combination of transformations. Moreover, the resulting library of graphs would be unbearably redundant, because every single paraphrase would need to be described in every single transformation graph's input.

2 Ambiguities

The chain of transformations that must be applied to a given elementary sentence in order to produce a given paraphrase constitutes the semantic analysis of the paraphrase. As I showed above, the meaning of the sentence *It is John who does not see a dog* is exactly represented by the formula:

*John sees a dog***[Neg] [Cleft-0]**

However, what happens where there is more than one way to produce a complex sentence from an elementary one? The sentence: *It was not the apple that was eaten by him?* seems to have two possible transformational analyses:

*John ate an apple***[Pron-0] [Passive] [Cleft-0] [Neg]**
*John ate an apple***[Passive] [Cleft-0] [Neg] [Pron-0]**

Are these two chains of transformation valid? If so, is this sentence really semantically ambiguous, or must these two analyses be unified? If so, what linguistic property can explain that some chains of transformations are to be ordered, whereas others might accept ordering variants? As transformation chains gets longer, the number of transformational paths that can be followed to analyze a single sentence will grow exponentially, and without any theoretical mean to distinguish between equivalent and different chains of transformations, the concept of transformational analysis itself becomes useless.[4]

[3] And we have not even discussed complex negations such as in *John wants not to come* or sentences that do not accept negations such as *John never comes* or *John does see the dog*.
[4] If a given sentence may have a large number of different analyses, what is the point of the analysis?

3 Theoretical Sentences

The self-imposed absolute need find a chain of transformation to link a given complex sentence to its initial elementary sentence is also responsible for another major drawback of transformational grammars: the so-called "theoretical sentences". Theoretical sentences are syntactically invalid sentences that have to be considered as valid in order to activate the transformational analysis of other (correct) sentences.

For instance, consider the French sentence: *La pomme est mangeable* [The apple is edible]. This sentence is traditionally linked to the elementary sentence *On mange la pomme* [One can eat the apple] via the following chain of transformations:

On mange la pomme **[pouvoir]**→ *On peut manger la pomme* **[Passif]**→ *La pomme peut être mangée* **[Able]**→ *La pomme est mangeable.*

Now, how can one analyze the sentence: *Luc est risible* [Luc is laughable] ? The same chain of transformations would be:

On rit de Luc **[pouvoir]**→ *On peut rire de Luc* **[Passif]**→ **Luc peut être rit* **[Able]**→ *Luc est risible.*

However, notice that the sentence **Luc peut être rit* [Luc can be laughed at] is incorrect. How then can one analyze the sentence *Luc est risible*? Either one allows users to add an invalid sentence to the grammar, so that an automatic system can process the complete chain of transformations from *On rit de Luc* to *Luc est risible*; or one forces users to design another chain of transformations that would bypass the invalid sentence, something like:

(a) *On rit de Luc* **[pouvoir]**→ *On peut rire de Luc* **[Passif-Able]**→ *Luc est risible.*

Or alternatively:

(b) *On rit de Luc* **[pouvoir-Passif-Able]**→ *Luc est risible.*

The first solution — which involves including a large number of so-called "theoretical sentences" in the grammar — has usually been chosen by most linguists who don't want to be bothered by mundane details such as the a-grammatical intermediary sentence[5]… The idea that the grammar of a language should contain a large number of a-grammatical sentences is difficult to defend.

The second solution (construct compound transformations to bypass invalid sentences) poses a theoretical problem: how do one decide which transformations will be compounded? How do one choose between the two transformational analyses (a) and (b)? And does it really matter? It makes no sense to give to a higher status to one or the other chain of transformations.

Moreover, this solution leads in practice to building a large number of compound transformations; in essence this solution is not different from the solution of constructing a pair of graphs for each pair of paraphrases; in effect it destroys the very idea of using elementary transformations.

The Automatic Generation of variants

[5] Indeed (Gross 1975) is a reaction to the general lack of interest for these "exceptions" and the LADL's project of constructing a systematic description of elementary sentences and their transformations.

4 Automatic generation of morphological variants

Generating a paraphrase from a given sentence is not very different from generating all the variants recognized by a given grammar. NooJ already can explore and generate all the paths of a morphological grammar. Consider the following morphological grammar **France** in Figure 4.

This grammar recognizes the word form *France* and associates it with the linguistic information "N+NomDePays" (it is a Noun and a name of a country). It also recognizes the word form *refranciser* and associates it with the information "V+FLX=AIDER+Re": it is a verb, conjugates according to the conjugation paradigm AIDER and it contains the repetition prefix.

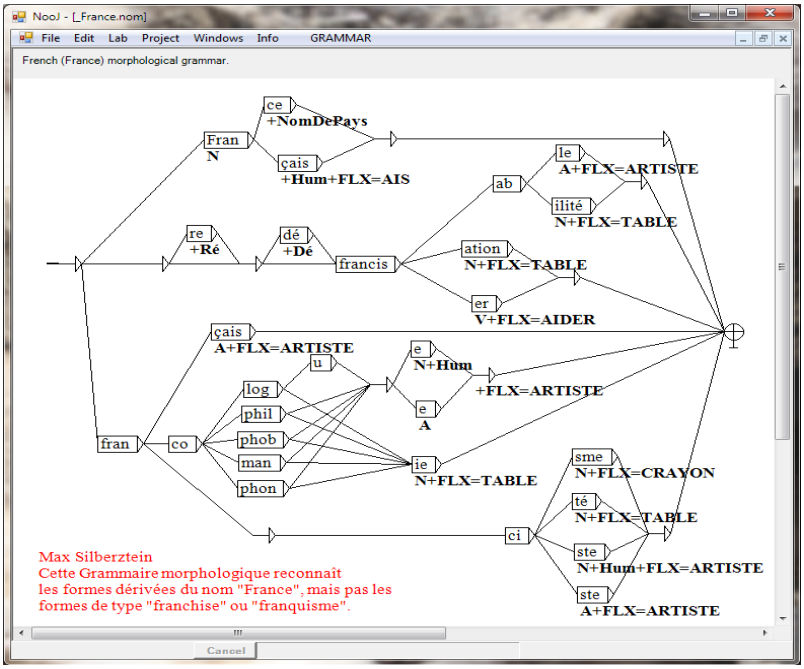


Figure 4. A morphological grammar

NooJ's simple command GRAMMAR > Generate Language can be used to automatically construct the dictionary that contains all the word forms recognized by the grammar, and associates each of the word forms with the corresponding grammar output.

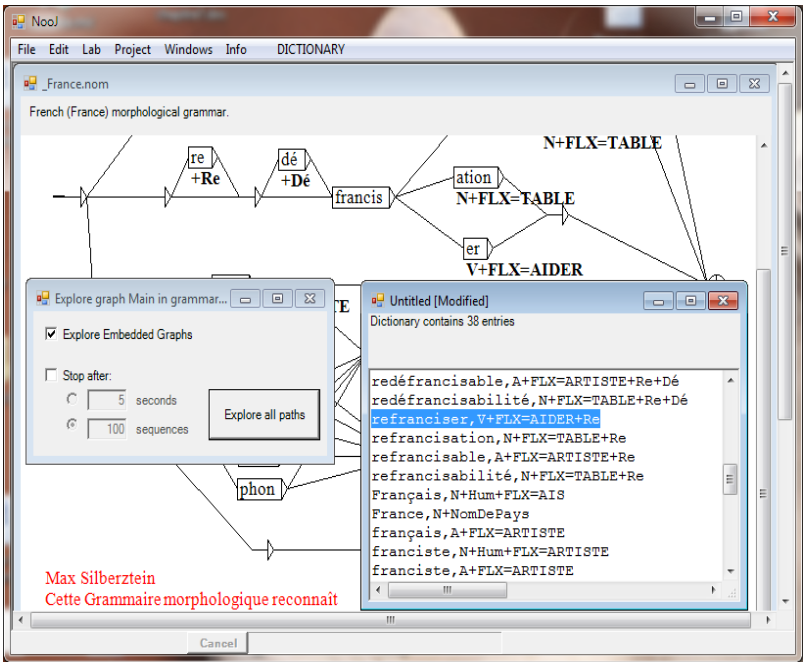


Figure 5. Generate all the word forms derived from *France*

Note that in the resulting dictionary, the linguistic information associated with each lexical entry corresponds in a sense to its linguistic analysis: analyzing the word form *refranciser* as a verb with the **+Re** (repetition) feature is similar to analyzing the sentence *John does not like apples* with the **+Neg** (negation) operator.

5 Automatic generation of syntactic paraphrases

If one constructs a syntactic grammar that recognizes all the paraphrases of a given elementary sentence, automatically generating all these paraphrases then becomes a matter of exploring the grammar. There is, however, a major difference between morphological and syntactic grammars : as opposed to morphological grammars that by design describe a specific family of derived forms, one cannot afford to build one grammar per elementary sentence: one needs to represent abstract structures such as N_0 *V* N_1 rather than plain sentences such as *John eats an apple*.

In order to produce the sentence *An apple is eaten by John* from the latter sentence, one needs to somehow process the sentence at the structure level:

$$N_0\ V\ N_1\ [\text{Passive}] \rightarrow N_1\ \textit{is V-pp by}\ N_0$$

and then manage variables so that N_1, *V-pp* and N_0 are correctly set respectively to *An apple*, *eaten* and *John*. And one wants the same exact grammar to produce the elementary sentence *John eats an apple* from the sentence *An apple is eaten by John*.

The solution is then (1) to construct a grammar that will recognize the elementary structure as well as all its paraphrases, (2) make sure the grammar is totally reversible, i.e. that it recognize every paraphrase (not only the elementary one), and (3) adapt NooJ's variable

management system and its morphological operators so that it can move variables' values around, and at the same time perform morphological operations if necessary.
The new grammar is very similar to any traditional NooJ grammar that would be used to recognize a sentence and its variants:

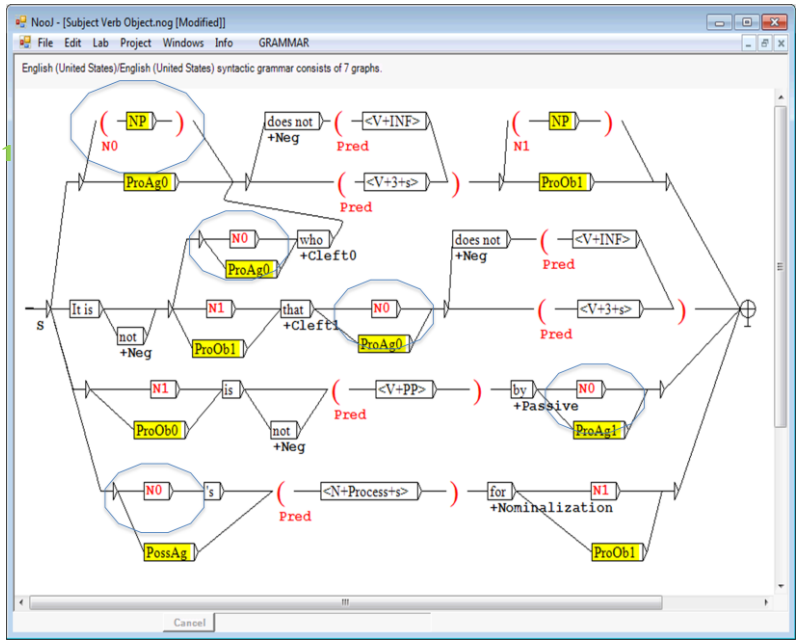


Figure 6. A syntactic grammar recognizes an elementary sentence and its variants

The grammar presented in Figure 6 is a "lighter" version of a syntactic grammar that could be used to perform a structural syntactic analysis of sentences, as described in (Silberztein 2010) and (Vietri 2010), which produce structural trees (I removed the informational structure in the interest of clarity).

6 Enhancements to syntactic grammars

The constraints and enhancements I have added to syntactic grammars so that they can be used as transformational grammars are:
— each subsequence that might be displaced as a whole by a transformation needs to be stored in a variable, here: $N0, $Pred and $N1[6];
— Any variable must be defined exactly once in a given grammar, because NooJ will unify the definition of the variable with every single instance of it in the grammar: this allows NooJ to move the content of the variable around, as well as to perform morphological operations on it.

[6] This very criterion is the one (Gross 1975) uses to characterize the syntactic constituents of a sentence.

For instance, in Figure 6, the variable $N0 has been set to the subgraph NP only once (see 1), and all other uses of $N0 in the same grammar refer to this same exact value: all the references $N0 in the grammar are linked to its definition (in the left top of the graph in Figure 6): "$(N0 :NP $)". This constraint allows any of this grammar's paths to match the corresponding paraphrases, rather than only the path at the top of the grammar.

Users might see this feature as a mere facility that makes grammars lighter and thus more readable (it is not necessary to redefine each of the occurrences of a given variable). In fact, however, solving references to a variable is crucial in order to allow NooJ to recognize not only elementary sentences (the path at the top of the graph) but also any paraphrase described in the grammar.

The way NooJ will compute variables' values has been enhanced so that the path $(Pred <V+PP> $) can produce the past-participle form *eaten* from the value of the $Pred variable *eats*, and reciprocally, that the path $(Pred <V+3+s> $) can produce the third person singular form *eats*, from the value of the variable $Pred *eaten*.

Figure 7 displays the automatic generation of all the paraphrases for one given sentence. The grammar produces the name of the transformations associated with each paraphrase, in other words, its transformational analysis. The list of features associated with each paraphrase always represents a transformation chain that would be performed on the elementary sentence, and not on the (possibly complex) sentence the grammar was applied to. In Figure 7, we see that the grammar, when given the complex sentence "Eva is not loved by Paul", has produced 34 paraphrases: each of these paraphrases is associated with the chain of transformations that links it to the elementary sentence: John loves Eva. For instance, the resulting sentence "Eva is loved by him" is associated with the chain S+Passive+Pro0, where Pro0 represents the pronominalization of the subject of the elementary sentence *Paul loves Eva* rather than the subject of the sentence *Eva is not loved by Paul*.

NooJ can be used to produce all the paraphrases of a given sentence (not necessarily an elementary one), or to produce only its paraphrases compatible with a certain transformational analysis. For instance, given the transformation chain "+Passive+Neg", NooJ will produce the four sentences:

she is loved by Paul, S+Pro1+Passive
she is loved by him, S+Pro1+Passive+Pro0
Eva is loved by Paul, S+Passive
Eva is loved by him, S+Passive+Pro0

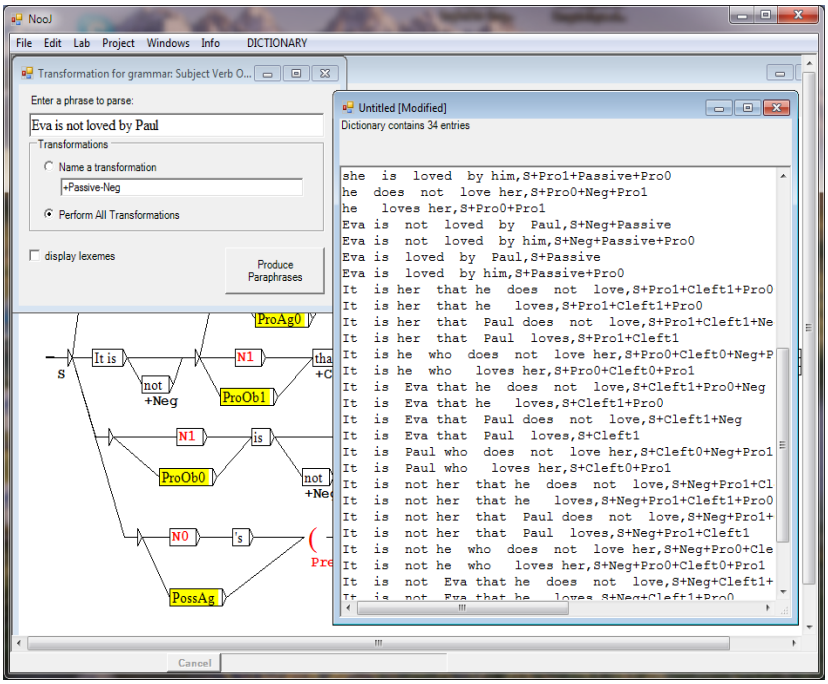


Figure 7. Produce all the paraphrases of the sentence Paul loves Eva

7 Conclusions

In this paper, I have presented an automatic transformational engine capable of producing any paraphrase from any given sentence (elementary or complex), as well as producing for any sentence its transformational analysis: its corresponding elementary sentence and the chain of transformations that link the elementary sentence to it.

As opposed to the traditional point of view on transformational grammars, the system presented here does not require linguists to implement a new level of linguistic description that would be different from the syntactic structural level: no need to implement specific transformational operators nor design a complex set of mechanisms to process chains of transformations.

NooJ's new transformational engine might be used in several Natural Language Processing applications, such as Question Answering: a question such as *Who ate the apple?* can be considered as a paraphrase of *John ate the apple*. If NooJ can link a question (typed in by a user) to any of its potential answers (occurring in a large corpus such as the WEB), it can answer the question automatically.

The functionalities presented here are a first attempt at the construction of a large-coverage formalization of elementary transformations: it will be necessary to test it on a large scale. In the future, I will need to design new grammars to represent more complex transformations, i.e. transformations that produce a complex sentence from more than one elementary sentence.

Keywords: NooJ, Syntactic Analysis, Transformational Analysis.

References

Ben Hamadou A., Piton O., Fehri H., 2010. Recognition and Arabic-French translation of named entities: case of the sport places. In the proceedings of the *NooJ 2009 International Conference and Workshop*. Sfax: Centre de Publication Universitaire, pp. 271-284.

Barreiro A, 2008. Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation. In Xavier Blanco & Max Silberztein (eds.), *Proceedings of the 2007 International NooJ Conference*. Newcastle: Cambridge Scholars Publishing, pp. 19-47.

Chomsky N., 1957. *Syntactic Structures*. The Hague: Mouton.

Fehri H., Haddar K., Ben Hamadou A., 2010. Integration of a transliteration process into an automatic translation system for named entities from Arabic to French. In the proceedings of the *NooJ 2009 International Conference and Workshop*. Sfax: Centre de Publication Universitaire, pp. 285-300.

Gross M. 1975. *Méthodes en syntaxe*. Paris: Hermann.

Harris Z. 1968. *Mathematical Structures of Language*. Englewood Cliffs, NJ: Interscience.

Silberztein M. 2010. Syntactic parsing with NooJ. In the proceedings of the *NooJ 2009 International Conference and Workshop*. Sfax: Centre de Publication Universitaire, pp. 177-190.

Vietri S. 2010. Building structural trees for frozen sentences. In the proceedings of the *NooJ 2009 International Conference and Workshop*. Sfax: Centre de Publication Universitaire, pp. 219-230.

Wu Mei. 2010. Integrating a dictionary of psychological verbs into a French-Chinese MT system. In the proceedings of the *NooJ 2009 International Conference and Workshop*. Sfax: Centre de Publication Universitaire, pp. 315-328.

Mary Astell's words in *A Serious Proposal to the Ladies* (part I), a lexicographic inquiry with NooJ

Hélène Pignot[(1)], Odile Piton[(2)]

[(1)(2)] *SAMM, University of Panthéon-Sorbonne*, Paris, France.

Abstract

In the following article we elected to study with NooJ the lexis of a 17th century text, Mary Astell's seminal essay, A Serious Proposal to the Ladies, part I, published in 1694. We first focused on the semantics to see how Astell builds her vindication of the female sex, which words she uses to sensitise women to their alienated condition and promote their education. Then we studied the morphology of the lexemes (which is different from contemporary English) used by the author, thanks to the NooJ tools we have devised for this purpose. NooJ has great functionalities for lexicographic work. Its commands and graphs prove to be most efficient in the spotting of archaic words or variants in spelling.

Introduction

In our previous articles, we have studied the singularities of 17th century English within the framework of a diachronic analysis thanks to syntactical and morphological graphs and thanks to the dictionaries we have compiled from a corpus that may be expanded overtime. Our early work was based on a limited corpus of English travel literature to Greece in the 17th century. This article deals with a late seventeenth century text written by a woman philosopher and essayist, Mary Astell (1666–1731), considered as one of the first English feminists. Astell wrote her essay at a time in English history when women were "the weaker vessel" and their main business in life was to charm and please men by their looks and submissiveness. In this essay we will see how NooJ can help us analyse Astell's rhetoric (what point of view does she adopt, does she speak in her own name, in the name of all women, what is her representation of men and women and their relationships in the text, what are the goals of education?). Then we will turn our attention to the morphology of words in the text and use NooJ commands and graphs to carry out a lexicographic inquiry into Astell's lexemes.

1 About the Author and Remarks on the Text

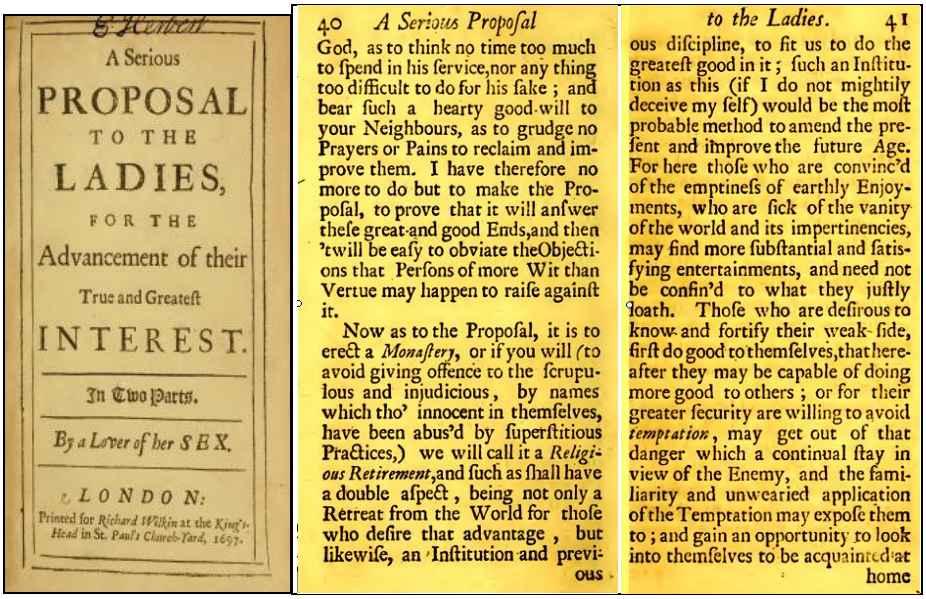
1. 1 Who was Mary Astell?

In *ASerious Proposal to the Ladies*, part I, first published anonymously in 1694, Mary Astell makes a rather odd suggestion (for the period!): women should withdraw from the world in a "blest abode" that she even dares to call a convent (in a country where all convents had been closed down). There they might enjoy each other's company, spend their time in study, good works and prayer (without forgetting daily church attendance). This place could also be a refuge for heiresses and unmarried women who need to escape the assiduities of adventurers and fortune seekers. Her book was widely discussed in her day and satirised by the moralist Richard Steele in the *Tatler* in 1709, calling Astell the leader

of "an order of Platonick Ladies" bent on celibacy and "resolv'd to join their Fortunes and erect a Nunnery" in Steele (1709).
According to the historian Antonia Fraser (1984, 404), what inspired Astell to write her proposal was the life of her friend Hortense Mancini, Duchesse de Mazarin, King Charles II's ex-mistress, who was the subject of numerous scandals, such as "the running away in Disguise with a Spruce Cavalier"! Astell was convinced her friend's "unhappy shipwreck" pointed out "the dangers of an ill Education and unequal marriage" (1700, 3). *ASerious Proposal to the Ladies*, part I, is signed "by a Lover of her Sex" and was first published anonymously in 1694 when the author was only 28. Part II (which will not be tackled here) is an application of Descartes' and Malebranche's philosophy to the education of women.

1. 2 From the Text to the Digital Text

OCR software could not be used for this text so we had to type the text preserving the original spelling and punctuation. NooJ does not keep the italics, hence the literary effects created by Astell's use of italics are lost. In 17th century printing, most nouns are capitalised. The last word (or its last syllable) on the page is repeated at the top of the next page: here "ous" and "home" in illustration 1.



A Serious
PROPOSAL
TO THE
LADIES,
FOR THE
Advancement of their
True and Greateſt
INTEREST.
In Two Parts.
By a Lover of her SEX.
LONDON:
Printed for *Richard Wilkin* at the *King's-Head* in St. *Paul's* Church-Yard, 1697.

40 *A Serious Propoſal*

God, as to think no time too much to ſpend in his ſervice, nor any thing too difficult to do for his ſake; and bear ſuch a hearty good-will to your Neighbours, as to grudge no Prayers or Pains to reclaim and improve them. I have therefore no more to do but to make the Propoſal, to prove that it will anſwer theſe great and good Ends, and then 'twill be eaſy to obviate the Objections that Perſons of more Wit than Vertue may happen to raiſe againſt it.

Now as to the Propoſal, it is to erect a *Monaſtery*, or if you will (to avoid giving offence to the ſcrupulous and injudicious, by names which tho' innocent in themſelves, have been abus'd by ſuperſtitious Practices,) we will call it a *Religious Retirement*, and ſuch as ſhall have a double aſpect, being not only a Retreat from the World for thoſe who deſire that advantage, but likewiſe, an Inſtitution and previ-

to the Ladies. 41

ous diſcipline, to fit us to do the greateſt good in it; ſuch an Inſtitution as this (if I do not mightily deceive my ſelf) would be the moſt probable method to amend the preſent and improve the future Age. For here thoſe who are convinc'd of the emptineſs of earthly Enjoyments, who are ſick of the vanity of the world and its impertinencies, may find more ſubſtantial and ſatisfying entertainments, and need not be confin'd to what they juſtly loath. Thoſe who are deſirous to know and fortify their weak ſide, firſt do good to themſelves, that hereafter they may be capable of doing more good to others; or for their greater ſecurity are willing to avoid *temptation*, may get out of that danger which a continual ſtay in view of the Enemy, and the familiarity and unwearied application of the Temptation may expoſe them to; and gain an opportunity to look into themſelves to be acquainted at

home

Figure 1. Frontispiece and Excerpt from *A Serious Proposal to the Ladies*

1.3 Astell's point of view

We first focused on the semantics, to examine how Astell builds her vindication of the female sex, which words she uses to sensitise women to their alienated condition and promote their education. As the object of this study is a literary text, it is interesting first to study point of view, which can be examined by focusing on the use of personal pronouns and possessive adjectives in Astell's text.

1.3.1 Personal pronouns

The study of personal pronouns highlight different aspects of Astell's rhetoric: she exhorts with the personal pronoun "you", denounces with "they", creates complicity with we, and seldom uses the first person "I".
The hortatory aspect of the text may be highlighted when we extract the sentences in which the pronoun "you" is used (locate pattern you): "How can you be content to be in the World like Tulips in a Garden, to make a fine shew and be good for nothing; have all your Glories set in the grave..." This functionality enabled us to extract expressions such as "I pray you", "I entreat you", "render you", "persuade you", I "would have you", "all that is required of you". The text is a protrepsis, an exhortation to women to change their lives and shake off the yoke of ignorance and superficiality.
The frequent use of the pronoun "we" is meant to encourage identification with the author and with a community of women sharing the same fate in a male-dominated society.
By contrast, the occurrences of "I, me, my" account for 4.56 % of pronouns. Astell does not use the pronouns "I" and "me" very much: these self-references (84) account for only 0.44% of the words of the text, which points to her self-effacement as an author, to be linked with that of the female author in the 17th century. Incidentally, it should be noted that her book is signed "by a Lover of her Sex", and not under her own name. In her preface to *Letters Concerning the Love of God* (1695,4), Astell says she wants "to slide gently through the world without so much as being seen or taken notice of."

Personal pronouns	occurrences	Total by group
Me/ my self / my / I	12 / 5 / 15 / 52	84
Thou / Thy / thine	0	0
She / her	129 / 183	129
He / him / his	18 / 24 / 24	631
We / us / our	141 / 88 / 144	373
You / your	125 / 81	206
They / them / their	148 / 111 / 159	418
Total	1841	

Table 1. Number of personal pronouns.

Such self-effacement was indeed common at the time as writing was not considered as an occupation befitting the female sex! It was only in the late 18th century that women were to establish themselves as authors and make a living as novelists. A talented novelist like Sarah Fielding, the sister of Henry Fielding, considered as one of the first 18th century female novelists, published all of her novels anonymously, and lived most of her life in strained circumstances. Indeed, she depended on the generosities of noble patrons who enabled her to publish her books by subscription. As the literary critic Vivien Jones points out: "*for women to write and publish at all was by definition a transgressive and potentially liberating act, a penetration of the forbidden public sphere, and the virulence with which fiction was attacked as a corrupting 'female' genre is telling evidence of its disruptive potential*" (1990, 12).

1.3.2 References to both sexes in the text

Now let us direct our attention to Astell's words. Is there anything about them that is provocative or subversive of traditional female roles and women's relationships with men? Is there any such thing as a feminine nature?
When we established the concordances for "man" and "woman", "husband" and "wife", masculine v. feminine and for the pronouns "they","we" and "you", we could extract key sentences like these ones, which show that all the supposed faults or vices that are attributed to women are not part of their nature, but result from a deficient education:

> *"if from our Infancy we are nurs'd up in Ignorance and Vanity; are taught to be Proud and Petulant, Delicate and Fantastick, Humorous and Inconstant, 'tis not strange that the ill effects of this conduct appear in all the future actions of our lives."*
>
> *"that Ignorance is the cause of most Feminine Vices, may be instanc'd in that Pride and Vanity which is usually imputed to us, and which I suppose -if thoroughly sifted, will appear to be some way or other, the rise and Original of all the rest. These, tho' very bad Weeds, are the product of a good Soil, they are nothing else but Generosity degenerated and corrupted. A desire to advance and perfect its Being, is planted by GOD in all Rational Natures, to excite them hereby to every worthy and becoming Action."*

Astell likes to use vegetal metaphors to make her point: uneducated women are compared to "*tulips in a garden*"[1] and education is likened to gardening. Women are rational (and not sentimental) creatures just like men: "*We value them too much, and our selves too little, if we place any part of our worth in their opinion; and do not think our selves capable of nobler Things than the pitiful Conquest of some worthless heart*"[2] (*Proposal*, 11) and they are trapped in the benign neglect in which men have purposely kept them: "*We're indeed oblig'd to them for their management...So that instead of inquiring why all Women are not wise and good, we have reason to wonder that there are any so. Were the men as much neglected, and as little care taken to cultivate and improve them, perhaps they wou'd be so far from surpassing those whom they now dispise, that they themselves wou'd sink into the greatest stupidity and brutality*" (*Proposal*, 15).
Women and men by custom are made to live in two separate spheres, respectively appearance and essence. Women are made by men creatures of appearance[3], frivolity and thoughtlessness, thanks to education they will have access to the realm of being and knowledge as God has given them –just as men— the faculty of reason: *"The Ladies, I'm sure, have no reason to dislike this Proposal, but I know not how the men will resent [=feel] it, to have their enclosure broken down, and Women invited to tast of that tree of Knowledge they have so long unjustly monopoliz'd. But they must excuse me, if I be as partial to my own Sex as they are to theirs, and think Women as capable of Learning as*

[1] "How can you be content to be in the World like Tulips in a Garden, to make a fine shew and be good for nothing" (Proposal, 9).
[2] And further down: "But I will not pretend to correct their Errors, who either are, or at least think themselves too wise to receive Instruction from a Womans Pen" .
[3] At one point in the text she calls dressing "that grand devourer" of time and energy, a remark mothers of teenage girls may still make today!

men are, and that it becomes them as well" (*Proposal,* 58). Incidentally, a clever, well-educated woman might also conveniently help conceal her companion's intellectual limitations[4]!

1.3.3 Astell's favourite words: the goals of women's education

In Astell's text (aside from grammatical words which are not significant here), the most commonly used lexemes are GOD (capitalized most of the time) and world. This may reflect a tension in the text between two conflicting demands for women in the 17th century: finding a place in the world but also accomplishing their spiritual destiny as God's creatures, endowed with reason just as males. The other most frequently used words in Astell's text are soul, virtue, nature, mind, knowledge and piety: these two notions are asserted throughout the text as the goals that women should pursue.

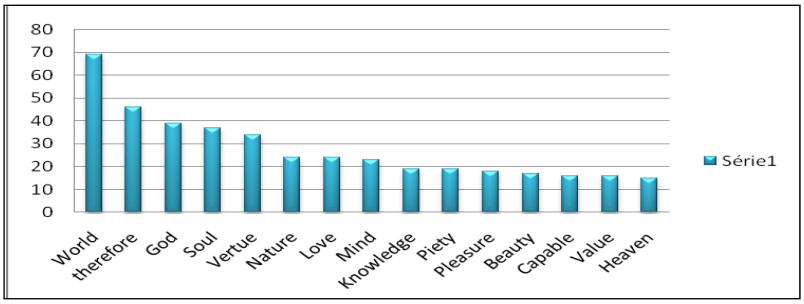


Table 2. Astell's 15 favourite lexemes

However, such lists and figures (even they reveal interesting frequent occurrences of certain words) can help literary analysis but not replace it. The absence of a word does not prove that the concept is not in the text; the writer may also use negation, circumlocution or euphemism. The words need to be put back into context (which can be shown by NooJ), as writers can resort to irony or humour, using words and meaning their exact opposites. The extraction of certain sentences from Astell's text highlights the author's frequent use of humour and irony.

Moreover, it must be pointed out that these listings do not provide any information on the sense relationships between individual lexemes. We are also aware that we have not studied collocations and frozen expressions in the text. However, within the scope of our work (recording the singularities of 17th century English), NooJ tools prove to be most efficient when the reader wants to study affixes in a 17th century text, as they help him locate archaic words or words whose prefixation, suffixation and spelling have changed over time.

2 Morphological study of Astell's words with NooJ

[4] "The only danger is that the Wife be more knowing than the Husband; but if she be 'tis his own fault, since he wants no opportunities of improvement; unless he be a natural Block-head, and then such an one will need a wise Woman to govern him, whose prudence will conceal it from publick Observation, and at once both cover and supply his defects."

After studying the length of words, we will record the morphological modifications we have observed and described with NooJ. Some lexemes are modified thanks to a punctuation mark such as an apostrophe (marking elision or contraction) or a hyphen. Other modifications include variations in affixes.

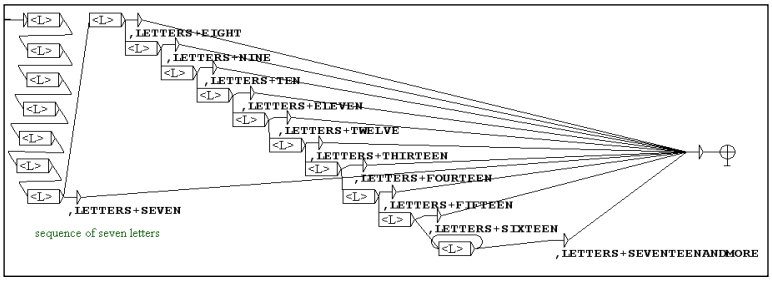


Figure 2. Morphological graph to locate long words

2.1 Long words in Astell's text

In English long words (as opposed to short Saxon words) are often learned words derived from Latin or Greek. A large number of Latin and Greek words were imported into the English language during the Renaissance. Astell's text is a literary work and we wondered if we might find the percentage of long words in her text thanks to NooJ. To do so, we created a graph that enables NooJ to locate them. We determined their number and the occurrences of each lexeme in a very simple manner. The graph below can tag the words in the text according to their length. Then to look for the occurrence of a N-letter word, we use the command <LETTERS+NUMBER>. Our results are presented in table 3:

LENGTH	WORDS	OCCURRENCES	OCC per WORD
SEVEN	515	1213	2.36
EIGHT	408	819	2.01
NINE	334	691	2.07
TEN	251	455	1.81
ELEVEN	144	248	1.73
TWELVE	79	121	1.53
THIRTEEN	37	68	1.84
FOURTEEN	16	27	1.69
FIFTEEN	10	15	1.5
SIXTEEN	1	1	1
Total	1795	3659	2.03

Table 3. Long words>6 letters

The total number of long words is 3659 out of 18,759 words in the text, i.e. a percentage of around 19.5%. This graph enables us to determine the percentage of long words and spot very long words (e.g. 16: uncharitableness) and some archaic words (e.g. 15: pragmaticalness). On average long words are used about twice (2.03).

2.2 Study of compounds and juxtaposed words

2.2.1 Hyphenated compounds

Many words were broken off and spelled as hyphenated compounds, instead of one single word: for instance fore-heads (17th century form) which is spelled foreheads in modern English. We have applied the following syntactical graph to be able to spot these forms. This graph highlights a systematic tendency to isolate the prefix from the root in 17th century spelling, thanks to the hyphenation (first path).

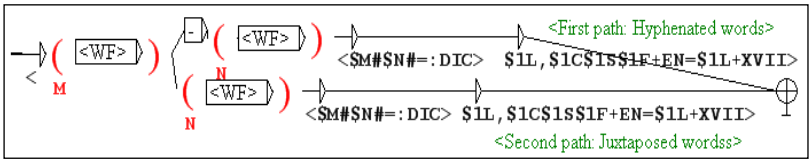


Figure 3. Syntactical graph for compounds

This produced 28 results or occurrences of the form word1-word2 (for example God-like for godlike or pre-ingage for preengage) recognised by NooJ.
During the search, NooJ also appeals to another graph which enables it to take into account the transformation of the prefix -in into -en or -en into -in (as we saw in a previous article, Piton and Pignot, 2008). Sometimes a fusing together of two hyphenated words may also need to be performed, as in "where-ever" which becomes "wherever". At our request NooJ was modified by Max Silberstein to permit this transformation.

NOUNS: 14		VERBS: 9		ADJECTIVES: 2		CONJ or ADV: 2	
Block-head	blockhead	me-thinks (impers verb) or methinks (arch): it seems to me		God-like	godlike	hence-forward	henceforward
church-yard	churchyard	out-do	outdo	over-careful	overcareful	where-ever	wherever
eye-sight	eyesight	out-weigh	outweigh				
fore-heads	foreheads	over-done	overdone				
good-will	goodwill	over-grown	overgrown				
Holy-day	holiday	over-rate	overrate				
ill-nature	illnature	over-run	overrun				
non-improvement	nonimprovement	over-stock'd	overstocked				
non-sense	nonsense	pre-ingage	preengage				
out-side	outside	(prefix in becomes en)					
pre-eminence	preeminence						
set-offs	setoffs						
well-fare	welfare						
well-wishers	wellwishers						

Table 4. From hyphenated words to concatenated words

2.2.2 Juxtaposed words

Some words result from the concatenation or fusing together of two lexemes, as in the case of personal pronouns (my self for myself). In 17th century English, these words were spelled as juxtaposed words instead of one single entity. We came across a few occurrences

of this type of words in Astell's text; the second path in illustration 3 enables us to look for them systematically. Besides personal pronouns such as ‘her self’, ‘our selves’ (whose spelling indicates that “self” was perceived as a noun), we have found 11 compounds: some body, every day, for ever, every one, any one, any thing, often times, like wise, to day, no body, back wardness. However, this link also produces much noise and funny results: *are a,area; be an,bean; direct or,director; not able,notable; yet I, yeti!* Some compounds are still hyphenated in contemporary English, while verbs (except when the prefix ends with the same letter as the first letter of the root) and adjectives no longer isolate the prefix thanks to the hyphenation.

2.3 Words with apostrophes

Frequence	329	2	90	356	1334
Character	'	‘	’	.	,

Table 5: Apostrophes in Astell's text

With a locate pattern of a single ', we have found 421 occurrences of apostrophes in the text. There are more apostrophes than full stops (356) and three times less than commas (1334). In 17th century typography, many words could be spelled with an apostrophe, which marked the elision of a letter, usually the mute e. Besides 118 past participles in ’d or ’t, some ’s and some familiar contractions, we have extracted other contractions that are not in the NooJ dictionary as they are not used in contemporary English. These contractions needed to be adequately described and added to our dictionary. Here are a few examples:
e’er, <ever,ADV> +UNAMB and 'twou'd, <it,it,PRO+3+n+s><would,V+PT+3+s> +UNAMB. Each entry describes the individual units in the contraction and we have tagged the entry as unambiguous (+UNAMB) when only one interpretation is possible. In the case of tho’t; the apostrophe is used to mark two elisions (an apocope and an aphaeresis): tho’= though and ‘t = it. Our dictionary (which includes Max’ Silberstein's 49 forms and ours) of contracted or elided forms has 90 entries. Many of these contractions are no longer used except in poetry (ex: cou'd : could; e'er: ever; ev'ry: every; heav'n: heaven). Any word containing an e mute could be elided in 17th century spelling (a graph could be made for this).

2.4 A Study of affixes in Astell's text

2.4.1 Inflectional affixes for adjectives, adverbs and verbs

In 17th century English any adverb or adjective whatever its length could form the comparative by adding the affix -er and the superlative by adding -est. The following graph may locate superlatives and comparatives; in Astell's text it can locate the archaic form worser (a double comparative of bad).

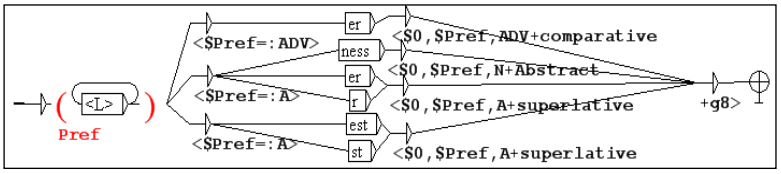


Figure 4. Recognition of the comparative and the superlative

Astell only uses the comparative in -er and the superlative in -est for short words.
Among grammatical affixes that are specific to 17th century English, mention should be made of the verb inflexions in -est and -eth (st, 2nd-person singular, -th or -st 3rd person singular) in the present and the preterit.
They may be spotted thanks to the graph below, and are recognised as valid verb forms when s or es is added to the root. It should be noted that there is only one occurrence of this form in our text ("profiteth") which dates from 1694.

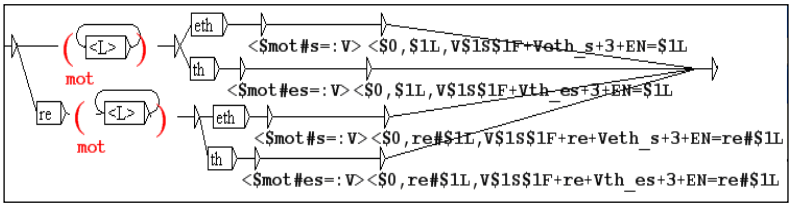


Figure 5. Morphological graph to recognise the third person present

We can also automatically recognise the absence of affix in the subjunctive in clauses introduced by a conjunction like lest (also spelled least): "least she only *change* the instance and *retain* the absurdity". The subjunctive was in full use as a thought-mood in 17th century English as we saw in a previous article (Piton and Pignot, 2008), not only after certain adjectives and verbs as in contemporary English, but in conditional, concessive and temporal clauses. The infinitive may be used for all persons.

2.4.2 Variations in affixes

In a previous article when we tackled our corpus of English travellers to Greece, we noticed variations in affixes (prefixes or suffixes). This new study is different as it simply uses simple existing NooJ commands to list prefixes and suffixes systematically and compare the word forms with their forms in contemporary English and spot the words that are archaic or have fallen into disuse. To do so, we have taken up the classification of affixes described by Tournier (2004). We decided not to specify the grammatical category (noun, verb, adjective) of the word form to make sure we could find all the results and sort them by grammatical category afterwards. We made a general inquiry into all the prefixes and suffixes mentioned by Tournier and only present the most significant results here.

Prefixes

As far as adjectival prefixes are concerned, Tournier's description includes quantitative prefixes (demi, semi, hemi, uni, ono, bi, di, diplo, ter/tri, quadri, quinque, penta, centi, hecto, ulti, pluri, poly), antonymic prefixes (a, in, un, non-, anti, dis) and spatial prefixes (epi, super, circum, peri, hypo, infra, sub, endo, intra). In Astell's text, there are no results for quantitative and spatial adjectives.
To locate prefixes thanks to NooJ, we need to use the following command <WF+MP="^prefix">, which can spot the variation of the prefix from "in" to "en" and vice versa for verbs: for the "en" prefix NooJ found 41 answers, such as encrease (increase in

CE); conversely for the in prefix we get 42 answers and five verbs could be found this way such as inforce, ingage, incourage (enforce etc. in contemporary English).
This method also allows us to locate words starting with "dis" with a different spelling or generally speaking long, obsolete words that are no longer used in contemporary English or with a different meaning. For the prefix -dis among 78 answers, it found the verb dispise, a spelling variant of despise in contemporary English which comes from the Latin *despicere*) and nouns such as disquisition (a formal enquiry, this word first appears in 1640 according to the OED), dispensatory (which means pharmacopoeia, a book in which medical substances and their properties are described). As the grammarian James Howell explains (1662, 10), e and i "supply one anothers place" and are used indifferently in English spelling at the time, just as in Spanish and Italian.

Suffixes

Dictionaries and grammars contain lists of prefixes. To locate suffixes in a text with NooJ, we need to use the command "locate pattern": <WF+MP="suffix$">. For NooJ, a suffix is a chain of characters, not a semantic entity. As far as nouns and verbs are concerned we observed variants that we have recapitulated further down. The main noun suffixes that we investigated are shown in table 6. The total number of archaic nouns or archaic forms for nouns located thanks to the study of suffixes is 26.

Suffix	Number of nouns	Number of occurrences	Nouns with an archaic meaning or spelling
tion/tions	98/17	171/25	commendation, concoction, disquisition
ity/ities	52/12	100/22	
ness/nesses	49/2	78/2	forwardness, pragmaticalness, back wardness, dearnesses
ment/ments	28/22	61/33	
ence/ences	25/7	47/12	
ance/ances	17/6	42/10	complyance
ure/ures	13/11	58/21	contexture
ency/encies	7/5	8/9	impertinency, indifferency, innocency, subserviency, displacencies
sion/sions	6/6	14/13	propensions
er/ers (human)	6/13	11/10	
or/ors	08/10	11/14	taylor, traytors
acy/acies	6/1	6/1	apostacy
ancy/ancies	4/1	5/2	incogitancy
ive/ives	4/3	5/4	preparative, persuasives
ary/aries	2/1	2/1	
ist/ists	2/0	2/0	
al/als	2/6	9/7	intellectuals, generals, temporals
ast/asts	1/0	1/0	antepast
yon/yons	0/1	0/1	cyons

Table 6. Archaic forms or meanings for nouns

For the suffix 'ness' we get 48 results, among which some archaic words: pragmaticalness (opinionatedness or stubbornness in CE), frowardness (disposition to disobedience and opposition), and one archaic form: back wardness spelled in two words that we added to our dictionary of 17th century English. The most productive suffixes seem to be -tion and -ness.
Here are other obsolete forms or meanings for words with the suffixes -ure, -tion, -cy, and -ce: contexture, disquisition, innocency, indifferency, subserviency, acquiesce, the verb "acquiesce in" having a specific meaning which is "remain at rest". The variation from -ence to -ency concerns nouns and when there are two forms of the same word the competing form in -ency tends to become archaic.

As regards adjectives and verbs, we systematically counted and studied adjectives in ous (117), al (91), able (61), ary (30), ly (26), ish (15), ate (15), less (15), ick (7), ical (6), ic (2), ose (0) and ist (0). We only observed 10 variations in spelling: woful (woeful); fantastick; heroick; flitting (fleeting), pityable (pitiable); changeable; improveable; unblameable; publick; vertuous (virtuous).
For verbs we found thirteen archaic forms: the verbs "disinteress", "attone", "glorifie" and "rectifie", six archaic spellings of the past participle (ex: "deprest, affixt"), and four gerunds: rejoycing (rejoicing), blustring, loosning and mouldring –in these three forms the mute e is elided.

2.5 Numerical assessment

If we do not apply the graphs and dictionaries we have created, NooJ finds 199 unknown words, such as improveable, improveing, incogitancy, incourage, inferiour, inforce, or ingage. The tools we have been developing with NooJ enabled us to correctly process 190 archaic words. After applying our graphs and dictionaries there were still 9 unknown forms (accrew, Aegypt, benumb, buz, cyons, meerly, mormo, shewn, viz) that we have added to our dictionary. In the case of "benumb" and "viz." these are not archaic forms or meanings but words that are missing from the sdic NooJ dictionary of contemporary English. Our systematic study of prefixes and suffixes enabled us to locate 49 archaic forms (nouns, verbs and adjectives).

Conclusions

From a semantic viewpoint, NooJ's functionalities (such as locate pattern) allow the reader to study the lexis of a text and single out its main lexical fields and recurrent words. Astell's essay is a very positive and optimistic advocacy of women's education, a proto-feminist text that has potent anti-male overtones as we saw in part I of this presentation. What motivates Astell in writing her text is above all a sincere wish to see women improve themselves thanks to education and fulfill themselves spiritually and intellectually.
When we read the text linearly, we counted 69 words out of 18,734 that have a specific meaning in 17th century English. NooJ could not locate every of them because their forms are in its dictionary, but could have other archaic meanings in 17th century English, for example "to close with a proposal" which means to accept it, "fantastick" which is synonymous with eccentric or quaint, "generals" (generalities), and "temporals" (temporal matters). We have added all these archaic meanings to our NooJ dictionary with a specific presentation which is meant to facilitate their recognition and listing. When the word varies only in spelling, the modern spelling is indicated next to the entry which reads: accrew,accrue,V+EN=accrue+Dic_EN_XVII+spelling.
When the meaning (not the form) of the word is different and archaic, we indicate it: displacency,N+EN=displeasure+Dic_EN_XVII+meaning. Each archaic meaning of a word necessitates a new entry.
All these NooJ tools will help us in a great but arduous work in progress, the creation of a dictionary of 17th century English which we would like to put at the disposal of the NooJ community.

Primary sources

Astell Mary. 1694. *A Serious Proposal to the Ladies: for the Advancement of their true and greatest Interest*. London.
Astell Mary. 1697. *A Serious Proposal to the Ladies, Parts I and II. Wherein a Method is offer'd for the Improvement of their Minds*. London.
Astell Mary. 1700. *Some Reflections Upon Marriage*. London.
Astell Mary. 2002. *A Serious Proposal to the Ladies, Parts I and II. Wherein a Method is offer'd for the Improvement of their Minds*, ed. P. Springborg. Broadview Literary Texts, Peterborough.
Astell Mary and John Norris. 1695. *Letters Concerning the Love of God*. London.
Howell James 1662. *A New English Grammar prescribing as certain rules as the English language will bear, for forreners to learn English*. London.
Miège Guy 1688. *The English Grammar or the Grounds and Genius of the English Tongue*. London.
Steele Richard 1709. *The Tatle*r, 23 June 1709.

References

Bauer, L. 1983. *English Word Formation*. Cambridge University Press, 1983.
Cottegnies, L. 2008. *Mary Astell et le féminisme en Angleterre*. ENS éditions, Lyon.
Fraser, A. 1984. *The Weaker Vessel: Woman's Lot in Seventeenth-Century England*. Phenix Press, London.
Jones, V. 1990. *Women in the 18th Century: Constructions of Femininity*. Routledge, London and New York.
Kolbrener W., Michelson M. 2007. *Mary Astell: Reason, Gender, Faith*. Ashgate, Newcastle.
Perry, R. 1986. *The Celebrated Mary Astell: An Early English Feminist*. Cambridge University Press, Cambridge, UK.
Pignot, H., Piton O. 2008. "Language processing of 17th Century British English with NooJ", *Workshop NooJ 2008*, Budapest, juin 2008
Piton O., Pignot, H. 2009. "Etude d'un corpus de textes de voyageurs anglais du 17esiècle, et aide à la transcription en anglais moderne", *6èmes Journées Internationales de Linguistique de Corpus*, Lorient.
Pignot H., Piton, O. 2010. "Mind your p's and q's: or the peregrinations of an apostrophe in 17th Century English", *Proceedings of the 2009 NooJ Conference*, Touzeur, Centre de Publication universitaire, 1-17.
Tournier, J. 2004. *Précis de lexicologie anglaise*. Ellipses, Paris.
Silberztein, M. 1993. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson.
Silberztein, M. 2005. "NooJ's dictionaries", *Proceedings of the 2nd Language & Technology Conference*, April 21-23, 2005, Poznán, Poland, Zygmunt Vetulani (ed.).
Silberztein, M. 2006. "NooJ's Linguistic Annotation Engine". In: Koeva, S., Maurel D., Silberztein M. (eds), *INTEX/NooJ pour le Traitement Automatique des Langues. Cahiers de la MSH Ledoux*. Presses Universitaires de Franche-Comté. 9-26.
Silberztein, M. 2007. "An Alternative Approach to Tagging". Invited paper. In: Proceedings of NLDB 2007. *LNCS series.* Springer. 1-11.

Sowal, M. 2008. "Mary Astell", *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/astell/, accessed 16 April 2010.

Building a Sanskrit module in NooJ: Basic resources

Vanja Štefanec
Faculty of Humanities and Social Sciences
University of Zagreb, Zagreb, Croatia

Abstract

Even though the first attempts to use the computer in Sanskrit studies were made already in the 1970s, the interest in computational processing of Sanskrit language increased only ten years ago. By now, a large number of linguistic tools have been developed and made available online, as well as the large corpora of digitalized Sanskrit texts. As to our knowledge, this is the first attempt to build resources for Sanskrit in NooJ.
Sanskrit was completely described in an amazing linguistic work Aṣṭādhyāyī composed by a famous Indian grammarian Pāṇini *in 5th century B.C. Idiom described in his grammar, known as Classical Sanskrit, basically remained unchanged till the present due to the fact that authors were consistent in following the grammatical rules.*
Sanskrit is a very complex language. It has a very rich and extremely regular inflectional and derivational morphology, loose syntax with almost free word order, very productive formation of compounds and high level of word sense ambiguity. But from the computational point of view, the most complex phenomenon is the sandhi – euphonic changes occurring between morphemes in a word (internal sandhi) or between words in a compound or a sentence (external sandhi), making thus the identification of words very difficult.
Although Sanskrit is a very important language for Indo-European comparative linguistics, this module should not be interesting only to linguists but to Sanskrit philologists as well. In the paper, we shall present some basic resources for Sanskrit module in NooJ, as well as give some remarks on few of the general problems in computational analysis of Sanskrit.

1 Introduction

The idea of building a computational model of a dead language might seem quite unusual, to say the least. But if we consider the enormous bulk of still unread texts written in that language and the fact that some of these texts were written by authors whose ingenuity is difficult to find a match to, the need for including the computational language processing in the research seems quite reasonable.
Researchers in Sanskrit philology, howsoever well versed in Sanskrit, often find themselves being unable to take all the relevant texts into consideration in their research. Our first, although pretty naïve, motivation for building this module was to try to overcome that problem.
To our knowledge, the first use of the NooJ framework for processing Sanskrit was described in Štefanec (2009) and this module is a result of continuation of that work.

Among other potential uses of this module, we could mention studies in Indo-European comparative linguistics or sociolinguistic studies dealing with modern spoken Sanskrit or Sanskrit loanwords in modern Indian languages.

2 The Sanskrit language

Sanskrit is an Indo-European language of the Indo-Iranian sub-family. When we speak about Sanskrit in a broader sense, we speak about two forms of language. The older is Vedic Sanskrit; spoken in the north-west of India in the period from around 1500 to 500 B.C. Vedic Sanskrit is the language of the Vedas. The younger form of the language, known as Classical Sanskrit, or simply Sanskrit, is the language described in the work *Aṣṭādhyāyī*, composed by a great Indian grammarian *Pāṇini* in the fifth or fourth century B.C. It is interesting to mention that *Pāṇini* described the language in formal rules using meta-language description, so Sanskrit became the first language in the world described by the principles of generative grammar. Compared to the Vedic Sanskrit, Classical Sanskrit underwent certain changes. There were minor changes in phonology, concerning mostly the innovations in sandhi. In morphology, inflectional paradigms were generalized and great number of grammatical forms became obsolete. Most of the innovations in Classical Sanskrit affected syntax and vocabulary. Although remarkably stable, vocabulary increased considerably by forming new words using grammatical rules without old restrictions, and, to much smaller extent, by borrowing words from Iranian and Dravidian languages, Greek and vernaculars. The major innovation in syntax was the increased tendency for using compounds, which lost any limitation concerning the number of elements (Burrow 2001: pp. 53-57). Indian tradition, however, never considered Vedic and Classical Sanskrit as different languages.

Classical Sanskrit is the language of classical Indian literature and culture. It has remained till the present the symbol of erudition and prestige. Sometimes it is referred to as "Indian Latin". Corpus of texts in classical Sanskrit is enormous, ranging from two great epics, *Mahābhārata* and *Rāmāyaṇa*, through great literary works of *kāvya* style, to the innumerous scientific texts, commentaries, commentaries on commentaries, and so on. Large number of them has not yet been read.

3 Lexical categories

Here we shall present several most important lexical categories in the Sanskrit language with their accompanying attributes in the form they will be described in the properties of the module. For the reason of limited space, we can not present them all, so we chose ones that are most representative.

3.1 Nouns and adjectives

In Sanskrit there are three grammatical genders, masculine, feminine and neuter, and three numbers, singular, dual and plural. Nouns and adjectives are inflected in eight cases: nominative, vocative, accusative, instrumental, dative, ablative, genitive and locative. To these eight, we shall add one more. That will be the crude form, form which the noun or adjective will take when serving as a first part of the compound. That form is often identical to the stem, but in the case of several classes of nouns and adjectives it will be different. For example, stems ending in *–in* and *–an* will lose the final consonant (cf. *dhanin* › *dhani-*, *rājan* › *rāja-*,etc.).

Figure 1 shows the frames of properties for nouns and adjectives.

```
N_Gender = m + f + n;
N_Nb = s + p + d;
```

```
N_Case = Nom + Voc + Acc + Ins + Dat + Abl + Gen + Loc + Cf;
A_Gender = m + f + n;
A_Nb = s + p + d;
A_Case = Nom + Voc + Acc + Ins + Dat + Abl + Gen + Loc + Cf;
```

Figure 1. Frames of properties for nouns and adjectives

3.2 Verbs

In Sanskrit, verbs are derived from verbal roots.
Sanskrit verbal system is fairly complex but very well structured. At the highest level, we can differentiate between 4 conjugations: primary, causative, desiderative and intensive. This refers to the way in which the conjugational stem will be formed. Type of conjugation models the basic meaning of the verbal root. Verbal tenses come at the second level. There are five actual tenses in Sanskrit and these are present (PR), imperfect (IMPF), perfect (PF), aorist (AOR) and future (FUT). To simplify the description of the verbal system, we shall stretch the definition of a tense and add to this list four more groups of "tenses". In the first group there are imperative (IMPR) and optative (OPT), which are the moods of present; benedictive (BEN), which is some kind of optative of the aorist; conditional (COND), which is some form of past future; injunctive (INJ), archaic mood from Vedic Sanskrit which can be found in Classical Sanskrit only in few syntactic constructions. Although this might seem a little bit confusing, there will not be any overlap between notions of tense and mood. In the second group we shall add two complex verbal forms formed by periphrasis, periphrastic perfect (PPF) and periphrastic future (PFUT). Forms that are not actual verbs but derived from them are included in the third group and these are absolutive (ABS), which is a verbal adverb, and infinitive (INF), which is a verbal noun fixed in accusative case. And, finally, the fourth group consists of forms that are not stand-alone words but bases for word formations; root (ROOT) and a verbal noun for formation of periphrastic perfect (PPFn). Category of grammatical voice comes at the third level. There are three grammatical voices: active, middle and passive. And at the end of the morphological description come the categories of grammatical person and number.
Proposed frame of properties is given in Figure 2.

```
V_Conj = prim + caus + inten + desid;
V_Tense = PR + IMPF + IMPR + OPT + FUT + PFUT + COND + PF + PPF +
PPFn +
                    + AOR + INJ + BEN + INF + ABS + ROOT;
V_Cond = act + med + pass;
V_Pers = 1 + 2 + 3;
V_Nb = s + p + d;
V_Prefix = prefixed;
```

Figure 2. Frame of properties for verbs

3.3 Participles

Participles are nominals derived from verbal stems and, as such, combine the attributes of both verbs and adjectives. Among the categories inherited from verbs there are still 4

conjugations, but here only four tenses: present (PR), future (FUT), perfect (PF) and past (PAST), and three voices. Categories inherited from nominals are 3 genders, 3 numbers and 8 + 1 cases.
Figure 3 shows the frame of properties for participles.

```
PRT_Conj = prim + caus + inten + desid;
PRT_Tense = PR + FUT + PF + PAST;
PRT_Cond = act + med + pass;
PRT_Gender = m + f + n;
PRT_Nb = s + p + d;
PRT_Case = Nom + Voc + Acc + Ins + Dat + Abl + Gen + Loc + Cf;
```

Figure 3. Frame of properties for participles

4 Used resources

The current version of the module is equipped with four large dictionaries of nouns, adjectives and numerals, verbs, participles and adverbs, which are recoded and reformatted to NooJ dictionary format from six morphological data banks derived from the Sanskrit Heritage Dictionary. Sanskrit Heritage Dictionary is developed by Gérard Huet within The Sanskrit Heritage Site and the framework of a computational linguistics platform for Sanskrit. These databanks are stored in XML format and contain more than a million inflected forms accompanied by lemma and morphological description.
The process of adapting these resources to fit the requirements of the NooJ framework included recoding the XML files into Unicode/UTF-8, change of transliteration standard from Velthius to IAST and, finally, reformatting the data and adapting the morphological description to the properties of the module.
Figure 4 shows three stages of that process on the example of one dictionary entry.

<f form="sa.msk.rti.h"><na><nom/><sg/><fem/></na><s stem="sa.msk.rti"/></f>	XML, ANSI; Velthius
<f form="saṃskṛtiḥ"><na><nom/><sg/><fem/></na><s stem="saṃskṛti"/></f>	XML, UTF-8; IAST
saṃskṛtiḥ,saṃskṛti,N+Nom+f+s	DIC, UTF-8; IAST

Figure 4. Illustration of the process of adapting external resources

5 Morphology

The amount of collected lexical stock might seem more than enough to satisfy all our needs regarding the vocabulary and morphology, but, due to highly productive word formation in Sanskrit, that is not the case. Sanskrit indeed has a very rich and extremely regular morphology, both inflectional and derivational, and from that point of view, we could say that it is ideal for computational processing. Many authors dealing with Sanskrit often emphasize certain morphological "transparency" of the language. By calling it

"transparent", they actually meant its regular word formation in which all word-forming constituents are recognizable in the resulting word, so the change of meaning is grammatically coded. For example, "quickly, instantly" can be said *kṣaṇeṇa*, which is actually instrumental singular of the noun *kṣaṇa* (= instant, moment), meaning literally "with a moment". "Bird" can be said *utpata* (= flying upwards), *pakṣin* (= winged), *pattrin* (= having feathers), *patat* (= flying), *taruśāyin* (= sleeping on trees), *nīḍin* (= having a nest), and so on… This "transparency" allows us to describe most of the word formation paradigms in derivational grammars in pretty much straight-forward way, reducing thus the need for building huge dictionaries. But, in the same time, we will have a high level of lexical ambiguity, which is one of the main disadvantages.

We shall now present several inflectional and derivational grammars that will enable the construction of additional dictionaries and recognition of derived forms.

5.1 Inflectional paradigms for nouns and adjectives

Inflectional paradigms for nouns and adjectives are defined in rules. Where possible, paradigms were named according to traditional grammars to simplify the construction of new dictionaries as much as possible. Fortunately, inflection of nouns and adjectives in Sanskrit is extremely regular which results in fairly small number of paradigms. In addition, just one look at the termination of a noun or adjective stem is enough for determining its inflectional paradigm.

The example of inflectional paradigm is shown in Figure 5.

```
saṃskṛti,N+FLX=MATI

MATI = <E>/Cf + ḥ/Nom+f+s + <B>e/Voc+f+s + m/Acc+f+s + <B>yā/Ins+f+s + <B>ye/Dat+f+s +
       <B>yai/Dat+f+s + <B>eḥ/Abl+f+s + <B>yāḥ/Abl+f+s + <B>eḥ/Gen+f+s +
       <B>yāḥ/Gen+f+s + <B>au/Loc+f+s + <B>yām/Loc+f+s + <B>ī/Nom+f+d +
       <B>ī/Voc+f+d + <B>ī/Acc+f+d + bhyām/Ins+f+d + bhyām/Dat+f+d + bhyām/Abl+f+d +
       <B>yoḥ/Gen+f+d + <B>yoḥ/Loc+f+d + <B>ayaḥ/Nom+f+p + <B>ayaḥ/Voc+f+p +
       <B>īḥ/Acc+f+p + bhiḥ/Ins+f+p + bhyaḥ/Dat+f+p + bhyaḥ/Abl+f+p + <B>īnām/Gen+f+p
       + ṣu/Loc+f+p;
```

Figure 5. Inflectional paradigm for feminine nouns ending in *–i*

5.2 Inflectional grammars for verbs and participles

Situation with verbs is somewhat different compared to nouns and adjectives. Everything that we have mentioned so far regarding the productive word formation mainly concerns nouns and adjectives. But verbs are the center of the Sanskrit lexical stock; verbal roots are the primary lexical units and a ground basis for primary word formation, and, as such, form a rather closed system. There are approximately 800 verbal roots which form the basis of the Sanskrit verbal system (Burrow 2001: p. 289) and our dictionary contains verbal and participle forms derived from nearly 650 roots. We believe that this can cover all the requirements within the module. However, we decided to provide some kind of a solution for entering new verbs if the need for that appears.

The second difference between the inflection of verbs and nouns is that the inflectional system of verbs is significantly more complex. For that reason, inflectional grammars for verbs and participles are defined graphically. Since the paradigms for the formation of

stems from roots would be impossible to generalize using computational devices included in the NooJ framework, verbs are supposed to be entered in the dictionary in their stem forms. That means that for every conjugation, theoretically six stems have to be entered to have all the forms covered: present active, present passive, future, aorist and perfect weak and strong stem. In addition, several forms cannot be derived from these stems, so they have to be entered separately: past passive participle, future passive participle, infinitive and absolutive. To avoid unnecessary overtagging, stems should be labeled as non-words in the dictionary. This solution might seem quite uneconomic, but there are actually two advantages of that approach. Roots, when forming stems, undergo different changes (apophony, reduplication, palatalization) and that process is difficult to generalize, so if we would insist on deriving the forms from a verbal root, we would have to provide a very large number of paradigms. Also, we wanted to use the same paradigms for deriving forms from denominative stems, in which case we cannot speak about the notion of root since they are derived from nominals. We have to stress once again that it is highly unlikely that there will be a need for any verb that already is not there in the basic dictionary.
In the example given in Figure 6, it is shown how the new verb should be entered in the dictionary to ensure that all the forms for primary conjugation are covered. Figure 6 also shows the inflection of the present active and middle forms for the *a*-stem.

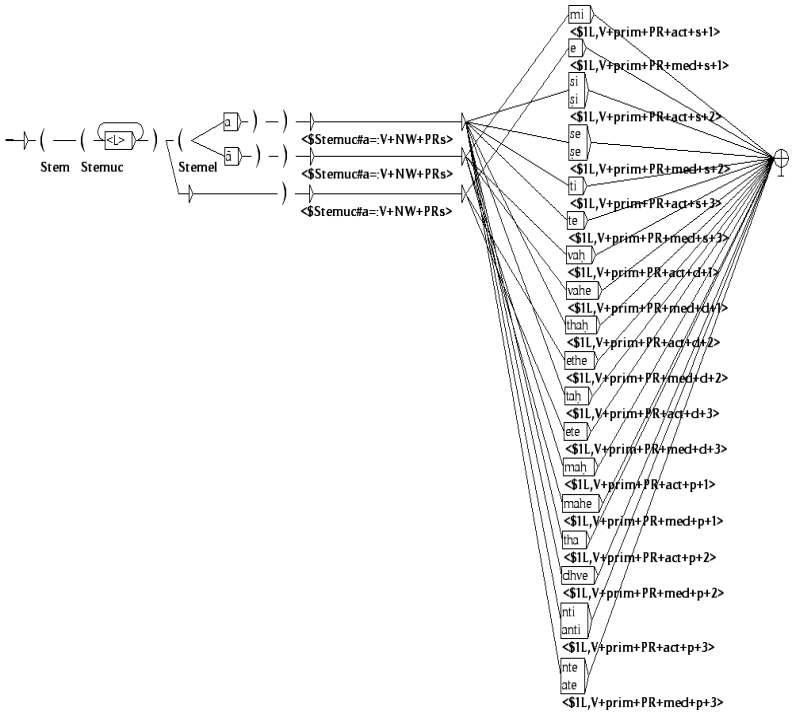


Figure 6. Example of entering new words in the dictionary and inflection of the present active and middle forms

biṭ,biṭ,V+NW+ROOT
beṭa,biṭ,V+NW+prim+act+PRs
biṭya,biṭ,V+NW+prim+pass+PRs
bibeṭ,biṭ,V+NW+prim+PTs
bibiṭ,biṭ,V+NW+prim+PTsw
abeṭiṣ,biṭ,V+NW+prim+AORs
beṭṭum,biṭ,V+prim+INF
biṭṭvā,biṭ,V+prim+ABS
biṭṭa,biṭ,PRT+prim+PAST+pass+FLX=KṚTA
beṭṭavya,biṭ,PRT+prim+FUT+pass+FLX=KṚTA

5.3 Derivational grammars

We have mentioned several times so far the derivational productivity of the Sanskrit language. Here we shall not deal with the primary formation of words, which is usually from verbal roots, since it is assumed that this part of lexical stock is mostly covered with the dictionary. We shall try to describe as many secondary word formations as we can, of which there is a significant number. Secondary word formation is done by means of suffixation; more specifically, by adding secondary derivational suffixes to words derived in primary formation. Euphonic changes occurring between the stem and the suffix do not pose a serious problem here (this will be explained in more detail in Section 6) but apophony is very difficult to deal with. Since we still have not found the adequate solution for this problem, derivations which require the change of the root vowel will stay undescribed for now. Such are, for example, *p**u**ruṣa* (= man) › *p**au**ruṣa* (= manly, human), *grīva* (= neck) › *gr**ai**vya* (= relating to the neck), etc.
Some of the derivations that can be described are formed by means of suffixes *–āyana* (formation of patronymics), *–in* (formation of adjectives), *–tva* (formation of neuter abstract nouns), *–tā* (formation of feminine abstract nouns), *–mat* and *–vat* (formation of adjectives in the meaning of a possessor of qualities), etc.
Example given in Figure 7 illustrates the derivational grammar for the formation of neuter abstract nouns in *–tva*. Cf. *am□ta* (= immortal) › *am□tatva* (= immortality), *tat* (= that) › *tattva* ("that-ness", truth, reality).

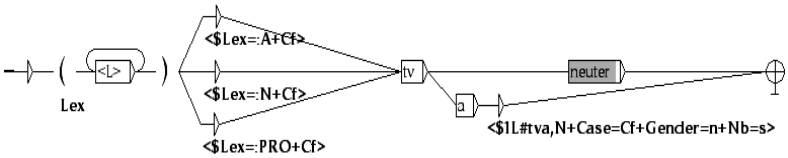


Figure 7. Derivation of neuter abstract nouns in *–tva*

6 Sandhi

Sandhi is an especially difficult phenomenon in Sanskrit from the point of view of computational analysis. Sandhi is the joint name for all euphonic changes occurring between morphemes in a word, i.e. internal sandhi, or words in a sentence, i.e. external sandhi. Sandhi occurs basically because of two phonologically motivated processes:

avoidance of hiatus and assimilation (Macdonell 2003: p. 10). Although such phonological changes caused by interaction of adjacent phonemes appear in probably all languages, in Sanskrit these changes are transferred into script. This means that in order to be able to identify the words from the verbatim transcribed spoken utterance, one should first undo the effect of external sandhi. We shall not deal here much with internal sandhi, because it does not pose a serious problem. Internal sandhi occurs within the morphemes in a word and we never have to analyze the word structure by resolving it. Moreover, applying sandhi is a simpler procedure than resolving it. Applying sandhi implies that the sandhi-forming constituents – or in the case of deriving new words by means of suffixation, one of them (specifically, the suffix) – and the place of occurring are already known, so there are not that many possible variants of the resulting euphonic change and all of them can be anticipated with the derivational grammar. However, when it comes to resolving sandhi, the degree of uncertainty is higher: neither place of occurring nor the constituents are known.

We shall briefly explain the mechanism of resolving external sandhi. Every word can be observed as a series of three sequences $\{(s_1)(s_2)(s_3)\}$. The first (s_1) is the variable part at the beginning that can change under the influence of a preceding word. Second (s_2) is the middle part that will not change under any circumstances and the third (s_3) is the final part that changes under the influence of the succeeding word. Having things put in this way, when three words are put together $\{(s_1)(s_2)(s_3)\}\{(s_1)(s_2)(s_3)\}\{(s_1)(s_2)(s_3)\}$, sandhi occurs in a following way $\{(s_1)(s_2)\{(\mathbf{s_{3\text{-}1}})\}(s_2)\{(\mathbf{s_{3\text{-}1}})\}(s_2)(s_3)\}$. Sequence (s_3) from the preceding word and sequence (s_1) from the succeeding will form new sequence $(s_{3\text{-}1})$, which can be seen as a unit that will always occur under the same circumstances. To illustrate, we shall take the sentence consisting of three words: *tat* (= that), *śrutvā* (= having heard) and *abravīt* (= said). First the words have to be split into aforesaid three sequences {(t)(a)(t)}{(ś)(rutv)(ā)}{(a)(bravī)(t)}. Then we apply the sandhi {(t)(a){(**cch**)}(rutv){(**ā**)}(bravī)(t)}. By applying sandhi, we see how this sentence would be written in Sanskrit: *tacchrutvābravīt* (= having heard that, he said). From this example we can generalize two rules: (1) *'-t'* followed by *'ś-'* gives *'-cch-'*, and (2) *'-ā'* followed by *'a-'* results in *'-ā-'*. There are many such rules ranging from simple (*'-m'* + *'a-'* › *'-ma-'*) to the most complex (*'-n'* + *'ch-'* › *'-ṃśch-'*).

Method of resolving external sandhi, which we are proposing within this module, is described in a morphological graph. It consists of finding those sequences in the string which can be the result of sandhi, checking the dictionary for the "corrected" forms of words in the constraint and, finally, adding the annotations. Since our grammar searches for around two hundred of such sequences, it is extremely large and cannot be presented here. So, as an example we shall compose a grammar which can split only the series of three words forming the *'-cch-'* and *'-ā-'* sandhi. Example is presented in Figure 8.

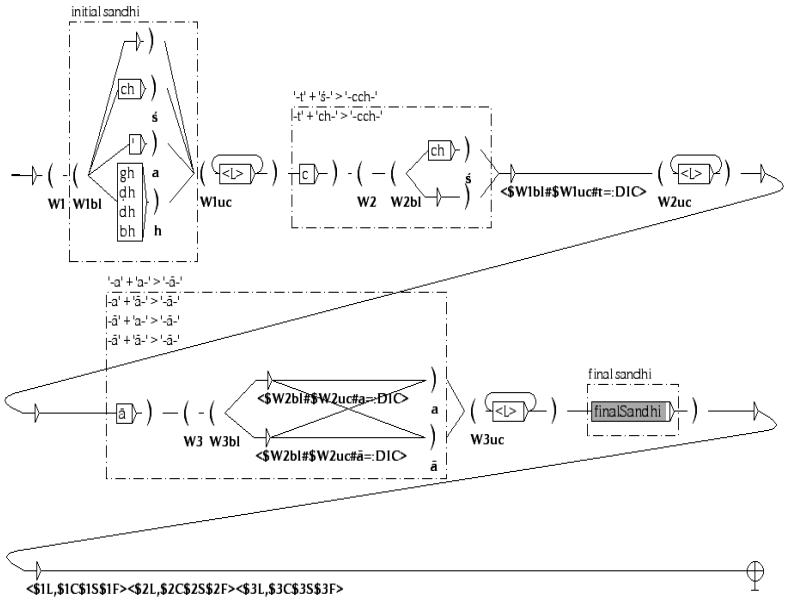


Figure 8. Example of resolving the *'-cch-'* and *'-ā-'* sandhi from a sequence consisting of three words

The main problem with this approach is that every syllable can be the result of some kind of sandhi and the number of words joined in a sequence can range from one to, theoretically, the number bigger than the number of syllables in that sequence (two syllables can contract to one). And, due to the fact that every of potential words have to be checked for, not just in the dictionary, but in every single derivational grammar, the process of splitting sandhi takes enormous amount of computational time. The second problem is, of course, ambiguity. Some sequences can be split in more than one way. Some of them will be syntactically completely impossible but on this level of analysis it cannot be determined. And what if we can have more than one syntactically correct splitting? These cases are not rare and might even not be accidental but might be the author's deliberate intention to add an additional (or sometimes even opposite) meaning. To sum up, not only that we have the problem of lexical ambiguity, we can also have more possible splittings and, finally, more possible readings of the same sequence. That leads to combinatorial explosion resulting in very large number of annotations that will have to be dealt with at the level of syntax.

The problem of sandhi in Sanskrit is in many respects similar to the problem of identifying words from a transcribed spoken utterance in speech-to-text systems so the proposed solution could be adapted for that purpose as well.

7 Syntax

Although we have not yet started dealing with syntax in this module, we shall present several phenomena from the domain of morphology which are described in syntactic graphs. One of them is periphrastic perfect, a complex tense formed by means of periphrasis. Figure 9 shows the proposed solution. As we can see, the first element carries

the lexical meaning, and the second element the morphological description. Although described in a syntactic graph, morphological annotation will be added to the form.

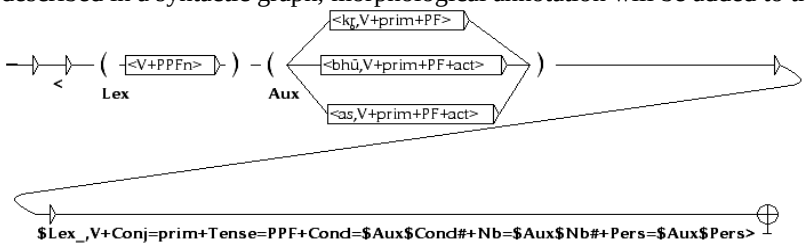


Figure 9. Recognition of periphrastic perfect

We shall see another use of syntactic graph for description of morphological phenomenon on the example of compounds. Although we agree with Speijer (1998: pp. 145-146) that compounds in Sanskrit belong in the domain of syntax, problem with which we shall be dealing here is purely morphological. As we have already mentioned, formation of compounds is extremely productive in Sanskrit and they have great syntactical importance (Macdonell 2003: p. 168). Language of compounds is really a language for itself and analyzing compounds consisting of up to 10 or more words presents one of the biggest challenges in dealing with Sanskrit.

A compound is always a combination of only two elements which can be either words or compounds themselves. According to Macdonell (2003: pp. 166-178), there are two types of compounds, verbal and nominal, and the latter can be further divided into 5 subtypes: *dvandva*, *dvigu*, *tatpuruṣa*, *karmadhāraya* and *bahuvrīhi*. In order to penetrate into the meaning of a compound, it is crucial to determine its type and subtype, which will give us information about the relation between its elements. At this stage, this is not yet possible, but it will definitely be one of the main goals in the future research. What we can do for now is to enclose all the words (or better yet, annotations assigned during the process of resolving sandhi) forming the compound and add the morphological description. As complex as it may be, compound has its clear syntactic role in a sentence and we can still work with it without going into its internal syntax. Figure 10 shows the grammar for recognition and annotation of nominal compounds. Compounds will be given syntactic annotation because there is no point in associating any lemma to it. As we can see from the grammar, we are searching for uninterrupted sequence of words in crude form, followed by a word in an inflected form. Morphological description is transferred from the final word to the whole compound, but its lexical category does not tell us anything about the lexical category of the compound.

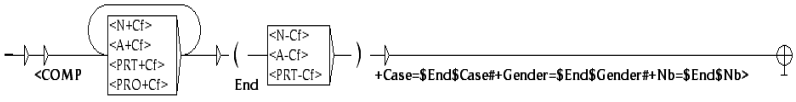


Figure 10. Recognition of nominal compounds

8 Textual forms

Thanks to several ongoing projects of digitalizing Sanskrit texts, we now have a very large corpus of texts in digital form. Although Sanskrit was throughout the history written in different scripts among which *devanāgarī* is most common, these texts are generally transcribed in Latin script. This is one of the reasons why the module was built using Latin script.

Digitalized versions of text differ according to editors' intervention. They sometimes reflect the way the text was written in the manuscript (whole sentences as uninterrupted sequences); sometimes the "spaces" are inserted where it is possible because of the sandhi and, in highly edited versions, all words, or even compounds can be split. In some versions, which then cannot be called plain text version but an analytic one, sandhi can also already be resolved. Moreover, since only two punctuation marks are used in Sanskrit, one for termination of a sentence and other for termination of a passage or verse, editors sometimes furnish the text with adequate punctuations for easier understanding.

Because of these varieties in textual forms, some phenomena had to be described both in morphological and syntactic grammars. One of these is aforementioned periphrastic perfect whose forms can then occur both as, for example, *"bodhayāmāsa"* and *"bodhayām āsa"* (= he awakened). In addition, if we go back to example in Figure 8 which illustrates the sandhi resolving method, we can see that even though the grammar is looking for complete sequences joined as a result of sandhi, the possibility that sandhi may occur both at the beginning and at the end has to be taken into account because we cannot assume the level of editor's intervention. For example, both *"tacchrutvā"* and *"tac chrutvā"* (= having heard that) can be found in the texts.

9 Conclusion and future work

In this paper, we have presented some general properties of a Sanskrit module for NooJ. All the resources that were built so far within the module are dealing with the morphology. Besides improving the present and building additional morphological resources, in our future work we shall definitely start dealing with the syntax. Moreover, our future work will also have to include finding the way of dealing with the ambiguity, which, as we have shown, will most likely pose a serious problem, as well as finding a suitable methodology for describing the internal syntax of the compounds.

Sanskrit studies is indeed a vast field of research and we hope that this module will be of help, not just in solving some of the unsolved questions, but also in avoiding the reluctance to open new ones.

Keywords: derivational morphology, inflectional morphology, NooJ, sandhi resolving, Sanskrit language

References

Burrow, T., 2001. *The Sanskrit language*, Delhi: Motilal Banarsidass Publishers.

Huet, G., 2009. Formal structure of Sanskrit text: Requirements analysis for a mechanical Sanskrit processor. In G. Huet, A. P. Kulkarni, & P. Scharf, eds. *Topics in Sanskrit Computational Linguistics*. Lecture Notes in Computer Science. Springer-Verlag. Available at: http://yquem.inria.fr/~huet/PUBLIC/Brown.pdf

Huet, G., 2010. *Sanskrit Heritage Dictionary: Morphological data banks vol. I - VI*, v246. Available at: http://sanskrit.inria.fr/DATA/XML/

Kale, M. R., 2002. *A higher Sanskrit grammar: for the use of schools and colleges*, Delhi: Motilal Banarsidass Publishers.

Macdonell, A. A., 2003. *A Sanskrit grammar for students*, Motilal Banarsidass Publishers.

Monier-Williams, M., 2008. *A Sanskrit-English dictionary*, digital edition, Universität zu Köln. Available at: http://www.sanskrit-lexicon.uni-koeln.de/monier/

Silberztein, M., 2003. NooJ manual. Available at: http://www.nooj4nlp.net. (223 pages)

Speijer, J. S., 1998. *Sanskrit syntax*, Delhi: Motilal Banarsidass Publishers.

Štefanec, V., 2009. Computational Methods for Detecting Formularity in the Works of Sanskrit Oral Epic Poetry. In *Proceedings of The First Middle European Student Indology Conference*. First Middle European Student Indology Conference. Zagreb. (in print)

Taraporewala, I. J. S., 1967. *Sanskrit syntax*, Delhi: Munshiram Manoharlal.

Greek in the age of corpora: Challenges and solutions

Dionysis Goutsos
Department of Linguistics, University of Athens, Greece
dgoutsos@phil.uoa.gr

Abstract

The paper offers a state-of-the-art description of corpus research on Greek, focusing on developments in corpus linguistics rather than computational linguistics. It refers to the specific characteristics of Greek that have had an effect on corpus research and outlines the main phases of development for Modern Greek corpora. It also presents the most important findings on the description of the Greek language deriving from corpora, with specific examples and references, and discusses the perspectives of corpus-related work on Greek.

1 Introduction

The hosting of the 10th NooJ conference in Greece offers an excellent opportunity to take stock of the main developments in previous and current corpus research on the Greek language. My perspective is that of the linguist who uses corpora and is interested in what corpora can reveal about language. In this view, I am following Hardie's (2009) distinction between computational linguistics, including language engineering and natural language processing, and corpus linguistics as two distinct fields that may overlap, but do not coincide.

Thus, this paper gives an overview of corpus linguistic work on Greek in order to complement the computational linguistic emphasis of this conference. I will first present the particular features of Greek that have influenced corpus development and analysis and will then refer to the development of Greek corpora for linguistic research. Next, I discuss the ways in which corpus analysis changes our view of the language on the basis of findings from several studies of Greek. Finally, the new challenges of Greek corpus linguistics are outlined with a view to suggesting further developments in the field.

2 Greek: Some peculiarities

A number of idiosyncratic features of Greek have been responsible for the slow rate of development of corpus linguistic research. First of all, Greek is a language with an especially long and complicated history. As Browning puts it, "since [the Homeric poems] Greek has enjoyed a continuous tradition down to the present day. Change there has certainly been. But there has been no break like that between Latin and the Romance languages. Ancient Greek is not a foreign language to the Greek of today as Anglo-Saxon is to the modern Englishman" (1983: vii). As a result, there have been multiple continuities and discontinuities in the history of Greek, leaving their traces in the language's structure and vocabulary. In addition, a complex sociolinguistic situation has emerged, that can broadly be characterized as Greek diglossia, which the language and its study have managed to overcome only in the late 1970s.

Without doubt, this is one of the reasons why Greek linguistics has shown an aversion to empiricism and only limited use of data in linguistic analysis. For instance, there is a gap of

some 50 years between the 1940s, when the first fully-fledged descriptions of Modern Greek appeared (Triandafyllidis, 1941; Tzartzanos, 1941-63), and the 1990s, when comparable, modern scientific descriptions were published (Holton et al., 1997; Clairis & Babiniotis, 1998-2004). The same is true for Modern Greek dictionaries, which only made an appearance in the late 1990s (Babiniotis, 1998; Idryma Manoli Triandafyllidi, 1998). Not surprisingly, these reference works were not based on corpus data.

Finally, the peculiarity of the Greek writing system, which differs from the standard Western alphabet, for which most computer applications were initially designed, created obstacles for the computational treatment of the language. Thus, the biggest part of the 1980s and 1990s was taken with the effort of the linguistic community to subvert the rigid ASCII code in order to accommodate the needs of the Greek user. The introduction of Unicode put an end to these problems and made further research redundant, but, meanwhile, a lot of time and resources was wasted on technicalities rather than the analysis of the language.[1]

3 The development of Greek corpora

Renouf (2007) distinguishes five stages in English language corpus evolution:

a) the 1960s-1970s, dominated by the one-million word Small Corpus (e.g. *LOB, Brown corpus*),

b) the 1980s, with the multi-million word Large Corpus or super-corpus (e.g. *Bank of English, BNC*),

c) the 1990s, with the 'Modern Diachronic' Corpus (e.g. *FLOB, Frown*)

d) 1998 onwards, during which the Web as corpus or cyber-corpus is introduced, and

e) 2005 onwards, expected to develop the Grid, i.e. a pathway to distributed corpora.

Because of the peculiarities mentioned in the previous section, among other reasons, Greek has been missing several of these stages. In particular, the first Greek corpora make an appearance in the late 1980s and early 1990s, when literary works are stored and analyzed by computational means (Philippides, 1981; 1986; 1988; Kyriazidis and Kazazis, 1992; Kyriazidis et al., 1992). In 1994 a survey finds that there are 15 small projects of collecting Greek data, but concludes that "corpora, if they are used in linguistic research at all, are not fully exploited" (Goutsos et al., 1994a: 215).

It is the 1990s and, especially, the 2000s which see the development of the two large corpora mainly used in Greek linguistics, the Hellenic National Corpus (HNC) and the Corpus of Greek Texts (CGT). HNC is a development of the Institute for Language and Speech Processing, currently including about 40 million words of mainly journalistic texts (Hatzigeorgiu et al., 2000).[2] CGT is a development of the Universities of Cyprus and Athens, including 30 million words from a wide range of spoken and written texts (Goutsos, 2003).[3] Goutsos (2010) compares the two corpora and argues that CGT fills the need for a representative and authoritative corpus of Greek, since HNC still includes a narrow range of text types, does not contain spoken data, has inadequate classification of texts and offers restricted availability.

[1] For a useful overview of problems and solutions with regard to corpora in the Greek alphabet, see King (1997).

[2] Available at: http://hnc.ilsp.gr/subcorpus.asp#

[3] Available at: http://www.sek.edu.gr

At the same time, a number of specialized corpora have started to make a late public appearance. These include the newspapers and school books corpora of the Greek Language Portal,[4] a biomedical corpus (Pantazara et al., 2007), as well as a Greek learner corpus and a thematic corpus, designed for learners at the University of Athens.[5]

It is clear that Greek is lagging behind other languages, in terms of both the size and the variety of corpora available for the description of the language. In Renouf's (2007) terms, it still lacks full 'super-corpora' and the dynamic, open-ended diachronic corpora available for English. However, research that has been based on the existing corpora has already borne fruit, as will be shown in the following section.

4 Corpus findings on Greek

There are several areas of Greek linguistics in which corpus-related research has produced a number of useful findings. These include the description of grammatical categories, phraseology, language variation, teaching applications, as well as the emergence of language norms and language change. The following presentation reviews the most important work in these areas, always from a corpus linguistic perspective.

4.1 Grammatical categories

Sinclair (1991) has pointed out that data-driven research, by avoiding predetermined linguistic categories, can identify facts about the grammar of a language which had previously been ignored. Thus, the study of corpora has pointed out the occurrence of the so-called shell nouns, general nouns that are used with several textual functions, including the encapsulation and labelling of a stretch of discourse. Koutsoulelou and Mikros (2004-2005) have studied the word *γεγονός* ('fact') in its use as a shell noun in the academic, journalistic and spoken sub-corpora of the CGT and found out its preference for the written mode, its collocations and phraseology, as well as its multiple functions and sub-functions. An extended study of all Greek shell nouns is still necessary in order to uncover similar patterns that will allow us to talk about a new sub-category of nouns in Greek.

Fragaki (2010a; 2010b) is a thorough investigation of Greek adjectives in an opinion articles sub-corpus of the CGT. Although the identification of adjectives follows pre-existing criteria, their classification is extensively corpus-driven, since it starts from evidence in the corpus. Thus, ten adjective categories are identified: classifying, descriptive, evaluative, deictic, relational, specializing, indefinite, colour, verbal and quantitative adjectives. Apart from important quantitative data, concerning e.g. the frequency of adjective categories and the relation between categories and their characteristics, the study also explores the evaluative and ideological role of adjectives in Greek discourse, pointing out that it is only certain adjective sub-categories that can take up these roles.

In all, corpus research has refined our knowledge of two basic grammatical categories of Greek; obviously, much more work is required before we have a full view of Greek grammar through corpora.

[4] Available at: http://www.greek-language.gr/greekLang/modern_greek/index.html

[5] Both available at: http://greekcorpora.isll.uoa.gr/gr/Default.aspx

4.2 Phraseology

Although the study of lexical collocations and phraseology seems to be ideally suited for corpus linguistic research, there have only been sporadic studies of Greek vocabulary (e.g. Goutsos et al., 1994b; Goutsos, 2009a).

An exception to this is the extended study of 3 to 5 word clusters (also known as lexical bundles or n-grams in the bibliography) in four Greek text types, spoken and academic texts, newspapers and fiction (Ferlas, 2011). Four different types of clusters are identified (basic, extended, variant and unique clusters), while the different functions they perform in discourse permits their categorization into the categories of stance, referential, text organizing, title, personal, grammatical and thematic clusters. What especially comes out in this research is the fact that Greek extensively draws on a number of word clusters such as *δεν μπορεί να* ('it cannot'), *δεν πρέπει να* ('it must not'), *θα μπορούσε να* ('it could'), *θα πρέπει να* ('it should') in order to indicate modality in discourse. In addition, the cross-linguistic comparison with English is made possible, pointing to similarities and differences between the two languages.

Again, more research from a corpus linguistic perspective is necessary in this area, in order to complement existing computational studies (e.g. Fragos et al., 2004).

4.3 Language variation

In a series of articles Mikros (1997; 2003; et al. 1996; et al. 2003; 2005, among else) systematically uses corpora in order to identify the parameters of phonological and morphological variation in Greek. This line of research has unearthed a host of interesting material on language variation and thus made possible an objective analysis of phenomena such as word couplets, which are due to the long history of diglossia (see section 2, above). Corpus linguistic methods are here combined with statistical and computational methodology in order to define basic characteristics of Greek texts. The findings of this research can thus be applied to such areas as the automatic identification of authorship, stylistic analysis etc.

Frantzi (2005) also uses statistical techniques in order to identify style features of political discourse. This is another area which is particularly interesting to explore, since it brings together the analysis of stylistic and ideological parameters of language variation in Greek.

Finally, the linguistic construction of gender identity has been studied in a couple of articles (Fragaki and Goutsos, 2005; Goutsos and Fragaki, 2009), which explore the meanings and collocations of gender-related nouns and adjectives in Greek (e.g. *άνδρας* 'man' vs. *γυναίκα* 'woman', *ανδρικός* 'male' vs. *γυναικείος* 'female'). This research has identified the ways in which gender asymmetry prevails in specific text types through patterns of nominal and adjectival use and their ideological implications. It is interesting to note that there have been only a few similar studies on other languages –mainly English– (see Goutsos and Fragaki, 2009: 319) and thus the area is offered for contrastive analysis.

4.4 Teaching applications

The main attempts to apply the findings of corpus linguistics in the teaching of Greek relate to the development of a specialized corpus for teaching Greek as a foreign language and a learner corpus, tagged for errors, both mentioned in section 3 above (see Iakovou et al., 2003). A similar project, aiming at the creation of a Corpus of Academic Greek Texts, is currently being developed at the Aristotle University of Thessaloniki, whereas a PhD dissertation on the basic academic vocabulary of Greek is currently written (Katsalirou, in prep.).

In addition, a first attempt at defining a basic vocabulary for Greek can be found in Goutsos (2006), which presents a number of basic nouns and verbs in Greek for both the CGT as a whole and sub-corpora of different text types, including academic texts, newspaper reports and opinion articles, legal-administrative and spoken texts.

4.5 Language norms and language change

One of the most important contributions of the corpus linguistic approach concerns the identification of language norms that cannot be reached at on the basis of intuition alone.

A case in point concerns the placement of connectives in Greek, which has been extensively studied in Goutsos (2009b). The category of connectives includes particles, discourse markers, sentence adverbials and other elements that are usually placed in the periphery of the clause and can have a crucial role in linking discourse rather than sentence parts. The area is notoriously difficult to divide into neat categories and, as a result, terms, both in Greek and other languages, proliferate, sometimes referring to the same phenomena. Corpus data can be invaluable in identifying frequent patterns and reaching generalizations about the linguistic behaviour of these elements.

In particular, the study of 1 million words of Greek from four sub-corpora of the CGT (academic texts, opinion articles, parliament speeches and TV and radio interviews) suggests that connectives show specific preferences for placement in particular clause positions. Table 1 below presents the figures in percentages for the occurrence of specific connectives at the beginning of the clause in the four sub-corpora.

	Academic	*Opinion articles*	*Parliament speeches*	*Interviews*
αντίθετα	65	89	56	88
άρα	60	56	86	97
επομένως	45	52	83	88
ευτυχώς	66	57	50	58
συνεπώς	55	76	97	100
εντούτοις	77	-	-	100
παρ' όλα αυτά	75	-	-	75
πρώτα-πρώτα	-	-	67	67
συμπερασματικά	100	100	-	-

Table 1. Connectives with preferred 1st clause position

As can be seen in Table 1, there are overwhelming tendencies for certain connectives such as adverbials of contrast (*αντίθετα, εντούτοις, παρ' όλα αυτά*) or conclusion (*άρα, επομένως, συνεπώς, συμπερασματικά*) and adverbials of stance (*ευτυχώς*) to occur in first clause position with frequencies that exceed half or even two thirds of their occurrences.

The importance of first clause position for connective elements has been stressed in the literature and has also been observed in several other languages (see Goutsos 2009b, for bibliography). Therefore, it is not surprising as such and can be accounted for on the basis of functional principles. What is more surprising is the tendency of several other Greek connectives to occur in second clause position, i.e. following the first clause constituent, as can be seen from the percentages of occurrence in Table 2.

	Academic	*Opinion articles*	*Parliament speeches*	*Interviews*
ακριβώς	50	45	47	48
άραγε	44	77	95	100
λοιπόν	80	88	87	65
όμως	74	88	75	60
πράγματι	40	38	52	45

Table 2. Connectives with preferred 2nd clause position

Corpus data suggest that this preference for second position is not accidental, since it concerns extremely frequent connective elements such as *λοιπόν* and *όμως* and holds for two thirds of their occurrences and across spoken and written text types, as can be seen in Table 2. In other words, it seems that second clause position has been conventionalized in Greek as the place for elements that indicate overall connectivity.

Again, several functional principles can be invoked to account for this (e.g. marking thematic position, rhythmic signalling etc.). However, what is most important is to find that connective elements in Greek have specific preferences for placement in the clause and that second clause position is reserved for some of these elements with surprising regularity across genres. These findings suggest that new norms have developed in Greek, about which little can be known without recourse to corpus data.

It is clear that the development of language norms is a predominantly diachronic phenomenon, which cannot be adequately studied in the absence of a diachronic corpus. Indirect evidence for language change can, however, be adduced, among others, through the study of new vocabulary that is introduced in Greek.

As is the case with other languages, computer terminology has been imported into Greek, mainly from English. There are three options for introducing new vocabulary in Greek, at least as far as the written mode is concerned: a) to use the foreign loan wholesale, i.e. in Latin characters (e.g. *computer, internet*), b) to transliterate the foreign loanword in Greek characters (e.g. *κομπιούτερ, ίντερνετ* or *ιντερνέτ*[6]) and c) to use a pre-existing Greek work (e.g. for computer *υπολογιστής* = calculator) or create a neologism by using pre-existing Greek morphemes (e.g. for internet *διαδίκτυο* from*δια*= inter and *δίκτυο*= network).

[6] The difference in transliteration concerns the placement of the accent according to the English or the French preference, respectively.

Although we cannot trace the history of the use of terms without a diachronic corpus (cf. Gorjanc 2006), a large corpus of Greek can offer evidence for the synchronic state of alternative uses. Figure 1 below presents the frequency of the terms used for *computer* in Greek, as found in the CGT. (A fourth option of the abbreviation *H/Y*, that is *ηλεκτρονικόςυπολογιστής* = electronic calculator, which is the full Greek equivalent for computer, has also been included).

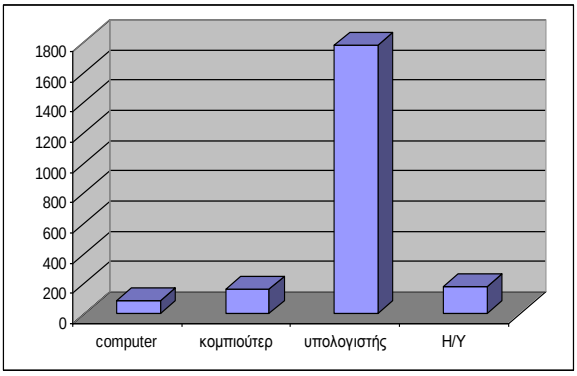


Figure 1. Frequency of alternative terms for 'computer' in CGT

As can be seen in Figure 1, the option that is overwhelmingly preferred in Greek is that of the Greek word, rather than the foreign term, either in Latin characters or transliterated. It is interesting to compare this data to figures from the Web; a Google search (January 2011) shows that the Greek word *υπολογιστής* is more than four times as frequent as the transliterated option κομπιούτερ (959.000 vs. 225.000 pages, respectively).

The numbers for the alternative terms for *internet* are shown in Figure 2.

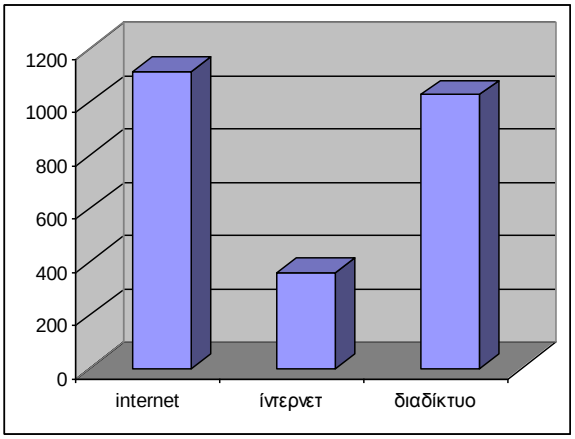


Figure 2. Frequency of alternative terms for 'internet' in CGT

Figure 2 suggests that the non-transliterated option is slightly more frequent than the Greek neologism and both are much more frequent than the transliterated alternative. The respective figures from a Google search favour the neologism *διαδίκτυο*, which occurs

almost three times as much as the transliterated option *ίντερνετ* (6.060.000 vs. 2.710.000 pages). This would confirm the trend found in the CGT in favour of the Greek neologism.

Although data from the Web offer updated evidence for current language use, conventional corpora like the CGT are invaluable in studying parameters that cannot be explored in the data offered by the internet. Thus, CGT can be used to study the frequency of the lemmas associated with the two options, as seen in Figure 3.

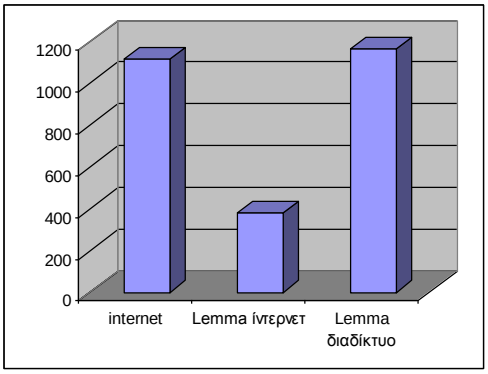


Figure 3. Frequency of alternative terms for the lemma 'internet' in CGT

Figure 3 compares the frequency of the 'non-Greek' *internet* with the frequency of the transliterated *ίντερνετ*, along with its derived nouns and adjectives (e.g. *ιντερνετικός*, *ιντερνετάκι*, even the plural noun*ιντερνέτια* etc) and that of the neologism διαδίκτυο, along with its derived nouns and adjectives (e.g.*διαδικτυακός*, *διαδικτυωμένος*, *διαδικτύωση*, *διαδικτυάκι* etc). It seems then that the lemma of the neologism is slightly higher in frequency than that of the transliterated option. This would suggest that the ease with which derived words can be formed in Greek affects the frequency and adoption of terms: the neologism thus offers an advantage over the other two options in being much easier to form derived words with.

In addition, a reference corpus like CGT is useful in comparing text types and thus identifying possible fields in which neologisms are to be found. In the case of the terms for *internet* in Greek, the frequencies presented above are split in Figure 4 according to text types.

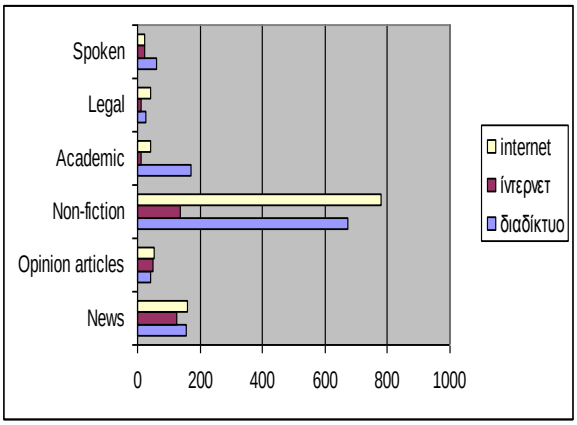


Figure 4. Distribution of frequencies of alternative terms for 'internet' in CGT text types

As shown in Figure 4, the non-transliterated option is more frequent in popularized, non-fiction texts, which is the text type in which all terms are much more frequently used. It also competes with the neologism in news, while the latter is much more frequent in academic texts, as well as spoken texts, although all terms are much less used in this text type.

In other words, non-fiction is the privileged area in which language change of this sort is expected to happen. In addition, while it is expected that academic texts would opt for the adapted Greek term, it is also interesting that spoken data confirm the preference for the neologism. This type of synchronic evidence can be crucial in determining the type and direction of potential language change.

5 Perspectives

The above discussion has made it clear that there are several areas in which corpus linguistic research on Greek is expected to develop in the future. First of all, there is an urgent need for compiling many more and more varied corpora with an emphasis on the diachronic study of Greek. This must include both the longer diachrony and recent language change. With respect to the former, a remaining challenge is to link Modern Greek corpora with corpora or databases for earlier phases of Greek such as ancient or Medieval Greek.[7] With respect to the latter, the challenge is to develop new, dynamic corpora, aimed at covering the decades of the 20th century before the 1990s and expand to the 21st century.

Secondly, the linguistic resources available on the Web can also be fruitfully explored to a larger extent than before, either through existing software such as Sketch Engine or WebCorp or through new methods of compiling corpora from the Web. This may include the compilation of comparable or parallel corpora that can be used in the analysis of Greek in contrast with other languages.

[7] For instance, see the projects available at: http://www.tlg.uci.edu and http://www.mml.cam.ac.uk/greek/grammarofmedievalgreek, respectively.

Thirdly, there is a need for increased interaction between corpus linguistic and computational linguistic methods. To this effect, the availability and standardization of NLP applications such as taggers, parsers etc. have to be improved.

Finally, there is much scope for the improvement of Greek language description with the use of corpus linguistic methods, including the investigation of grammatical categories and specialized phraseology, as well as through the study of new text types and genres of Greek. The final aim would be to produce new grammars and dictionaries[8] that would be based on less intuitive and more accurate empirical data.

Acknowledgments

I would like to thank the organizers of NooJ 2010, and especially Zoe Gavriilidou, for their kind invitation to give this plenary lecture, for their hospitality and for giving me the opportunity to meet the NooJ family. Many thanks to Philip King for his always valuable comments.

References

Babiniotis, G. 1998. *Λεξικό της Νέας Ελληνικής Γλώσσας.* [Dictionary of Modern Greek]. Kentro Lexicologias, Athens.

Browning, R. 1983. *Medieval and Modern Greek.* Cambridge University Press, Cambridge.

Clairis, C. and Babiniotis, G. 1998-2004. *Γραμματική της Νέας Ελληνικής. Δομολειτουργική-Επικοινωνιακή.*[Modern Greek Grammar. Structural-Functional-Communicative]. Ellinika Grammata, Athens.

Ferlas, H. 2011. *Ο προκατασκευασμένος λόγος στα Ελληνικά και Αγγλικά. Μια μελέτη βασισμένη σε σώματα κειμένων με προεκτάσεις στη διδασκαλία της γλώσσας.*[Prefabricated discourse in Greek and English. A corpus-based study with implications for language teaching]. PhD dissertation, University of Athens.

Fragaki, G. 2010a. *Ο αξιολογικός ρόλος του επιθέτου και η χρήση του ως δείκτη ιδεολογίας: Μελέτη βασισμένη σε σώματα κειμένων δημοσιογραφικού λόγου* [The evaluative role of the adjective and its use as a marker of ideology: A study based on journalistic corpora]. PhD dissertation, University of Athens.

Fragaki, G. 2010b. A corpus-based categorization of Greek adjectives. *Proceedings of the 5th Corpus Linguistics Conference. 21-23 July 2009. University of Liverpool.* Available at: http://ucrel.lancs.ac.uk/publications/CL2009/.

Fragaki, G. and Goutsos, D. 2005. Gender adjectives and identity construction in Greek corpora. *Proceedings of the 7th International Conference on Greek Linguistics, University of York, 8-10 September 2005*. Available at: http://83.212.19.218/icgl7/Fragaki-et-al.pdf.

Fragos, K., Maistros, I. and Skourlas, C. 2004. Extracting collocations in Modern Greek language. *Proceedings of the 1st International Conference on Natural Language Understanding and Cognitive Science, Porto, Portugal, 13-14 April 2004.* Available at: http:// glotta.ntua.gr/nlp_lab/Fraggos/files/DiCofinal.pdf

[8] For the potential contribution of corpora to Greek lexicography, see Goutsos & Fragaki (forthcoming).

Frantzi, K. 2005. Γλωσσικά και μη χαρακτηριστικά των προκηρύξεων της 17Ν. [Linguistic and non-linguistic features of terrorist manifestoes]. *Studies in Greek Linguistics* 25: 639–650.

Gorjanc, V. 2006. Tracking lexical changes in the reference corpus of Slovene texts. In Andrew Wilson, Dawn Archer and Paul Rayson (eds) *Corpus Linguistics Around the World.* Rodopi, Amsterdam, 91-100.

Goutsos, D. 2003. Σώμα Ελληνικών Κειμένων: Σχεδιασμός και υλοποίηση. [Corpus of Greek Texts: Design and implementation]. *Proceedings of the 6th International Conference on Greek Linguistics, University of Crete, 18-21 September 2003.* Available at the webpage: http://www.philology.uoc.gr/conferences/6thICGL/gr.htm.

Goutsos, D. 2006. Ανάπτυξη λεξιλογίου. Από το βασικό στο προχωρημένο επίπεδο. [Vocabularydevelopmentfromthebasictotheadvancelevel]. *Η ελληνική ως ξένη γλώσσα: Από τις λέξεις στα κείμενα.* Patakis, Athens, 13-92.

Goutsos, D. 2009a. «Λόγος να γίνεται»: Μεταγλωσσική φρασεολογία στο λόγο των πολιτικών του κοινοβουλίου. [Metalinguistic phraseology in Parliament discourse]. In Eleni Karamalengou and Eugenia Makrygianni (eds). *Αντιφίλησις. Studies on Classical, Byzantine and Modern Greek Literature and Culture. In Honour of John-Theophanes A. Papademetriou*. Franz Steiner, Stuttgart, 638-647.

Goutsos, D. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora* 5 (1): 29-44.

Goutsos, D. and Fragaki, G. 2009. Lexical choices of gender identity in Greek genres: The view from corpora. *Pragmatics* 19 (3): 317-340.

Goutsos, D. and Fragaki, G. Forthcoming. Λεξικά και σώματα κειμένων. [Dictionaries and corpora]. In Γιώργος Ξυδόπουλος, Άγις Οικονομίδης and Γιώργος Τράπαλης (eds). *Εισαγωγή στη λεξικογραφία*. [Introduction to Lexicography]. Patakis, Athens.

Goutsos, D., King, P. and Hatzidaki, O. 1994b. A corpus-based approach to Modern Greek language research and teaching. In Irene Philippaki-Warburton, Katerina Nicolaidis and Sifianou Maria (eds) *Themes in Greek Linguistics: Papers from the First International Conference on Greek Linguistics. Reading, September 1993*. John Benjamins, Amsterdam/Philadelphia, 507-513.

Goutsos, D., King, P. and Hatzidaki, R. 1994a. Towards a Corpus of Spoken Modern Greek. *Literary and Linguistic Computing* 9 (3): 215-223.

Goutsos, G. 2009b. Μόρια, δείκτες λόγου και κειμενικά επιρρήματα: Η οριοθέτηση των γλωσσικών κατηγοριών με τη χρήση ΗΣΚ. [Particles, discourse markers and text adverbs: The definition of linguistic categories through the use of corpora] *Proceedings of the 8th International Conference on Greek Linguistics, University of Ioannina, 30 August-2 September 2009*, 754-768. Available at: http://www.linguist-uoi.gr/cd_web/ case2.html.

Hardie, A. 2009. Corpus linguistics and the languages of South Asia: Some current research directions. In Paul Baker (ed.) *Contemporary Corpus Linguistics*. Continuum, London, 262-288.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H. and Demiros, I. 2000. Design and implementation of the online ISLP corpus. *Proceedings of the LREC 2000 Conference*, Athens, 1737-1742.

Holton, D., Mackridge, P and Philippaki-Warburton, I.. 1997. *Greek. A Comprehensive Grammar of the Modern Language.* Routledge, London.

Iakovou, M., Markopoulos, M and Mikros, G. 2003. Θεματοποιημένο Βασικό Λεξιλόγιο μέσω ΗΣΚ: Πρακτική εφαρμογή στη διδασκαλία της Νέας Ελληνικής ως ξένης γλώσσας. [Thematic basic vocabulary through corpora. A practical application to the teaching of Modern Greek as a foreign language]. *Proceedings of the 6th International Conference on Greek Linguistics, University of Crete, 18-21 September 2003.* Availableatthe webpage: http://www.philology.uoc.gr/ conferences/6thICGL/gr.htm.

Idryma Manoli Triandafyllidi. 1998. *Λεξικό της κοινής νεοελληνικής.* [Dictionary of Standard Modern Greek]. Institute of Modern Greek Studies, Thessaloniki.

Katsalirou, A. Inpreparation. *Το λεξιλόγιο για γενικούς ακαδημαϊκούς σκοπούς στη διδακτική της νέας ελληνικής ως ξένης γλώσσας.* [Vocabulary for general academic purposes in the teaching of Modern Greek as a foreign language]. PhD dissertation, Aristotle University of Thessaloniki.

King, P. 1997. Creating and processing corpora in Greek and Cyrillic alphabets on the personal computer. In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds). *Teaching and Language Corpora*. Longman, London, 277-291.

Koutsoulelou, S. and Mikros, G. 2004-2005. Το «γεγονός» ως ουσιαστικό κέλυφος. Χρήση και λειτουργία σε ηλεκτρονικά σώματα κειμένων της Ελληνικής. [Γεγονός as a shell noun. Use and function in Greek electronic corpora]. *Glossologia* 16: 65-95.

Kyriazidis, N. and Kazazis, I., K. 1992. *Τα ελληνικά του Μακρυγιάννη με τον υπολογιστή*. [Makriyannis' Greek in the computer]. Papazisis, Athens.

Kyriazidis, N., Kazazis, I., N. and Brehier, J. 1992. *Το λεξιλόγιο του Μακρυγιάννη.* [Makriyannis' vocabulary]. Athens.

Mikros, G. 1997. Radio news and phonetic variation in Modern Greek. *Greek Linguistics '95. Proceedings of the 2nd International Conference on Greek Linguistics 1995*, I, 35-44.

Mikros, G. 2003. Στατιστικές προσεγγίσεις στην αυτόματη κατηγοριοποίηση κειμένων της Νέας Ελληνικής: Μια πιλοτική αξιολόγηση υφομετρικών δεικτών και στατιστικών μεθόδων. [Statistical approaches to the automatic classification of Modern Greek texts. A pilot evaluation of stylometric indexes and statistical methods]. *Proceedings of the 6th International Conference on Greek Linguistics, University of Crete, 18-21 September 2003.* Availableatthe webpage: http://www. philology.uoc.gr/conferences/6thICGL/gr.htm.

Mikros, G., Gavriilidou, M., Lambropoulou, P. and Doukas, D. 1996. Χθες ή χτες; Μια ποσοτική μελέτη φωνητικών και μορφολογικών στοιχείων σε κείμενα της Νέας Ελληνικής. [A quantitative study of phonetic and morphological features in Modern Greek texts]. *Studies in Greek Linguistics* 16: 645-656.

Mikros, G., Hatzigeorgiu, N. and Carayannis, G. 2003. Βασικά ποσοτικά μεγέθη στην γραπτή Νέα Ελληνική γλώσσα: η αξιοποίηση του ΕΘΕΓ στην ελληνική ποσοτική γλωσσολογία. [Basic quantitative measures in written Modern Greek. The exploitation of HNC in Greek quantitative linguistics. *Proceedings of Workshop on Text processing for Modern Greek: From Symbolic to Statistical Approaches", Rethymno, 20 September 2003*, 23-37.

Mikros, G., Hatzigeorgiu, N. and Carayannis, G. 2005. Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics* 12 (2-3): 167-184.

Pantazara, M., Mantzari, E., Vagelatos, A., Kalamara, C. and Iordanidou, A. 2007. Development of a Greek biomedical corpus. Paper given at the 11th Panhellenic Conference on Informatics (PCI 2007), Patras, Greece. Available at: http://www.iatrolexi.gr/vagelat/Iatrolexi-corpus.pdf.

Philippides, D. 1981. Computers and Modern Greek. *Mantatoforos* 17: 5-13.

Philippides, D. 1986. *The Sacrifice of Abraham on the Computer.* Hermes Press, Athens.

Philippides, D. 1988. Literary detection in the *Erotokritos* and *The Sacrifice of Abraham*. *Literary and Linguistic Computing* 3: 1-11.

Renouf, A. 2007. Corpus development 25 years on: from super-corpus to cyber-corpus. In Roberta Facchinetti (ed.) *Corpus Linguistics 25 Years on.* Rodopi, Amsterdam 27-49.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Triandafyllidis, M. 1941. *Νεοελληνική γραμματική (της δημοτικής)* [Modern Greek Grammar (of dimotiki)]. Organismos Ekdoseos Sxolikon Vivlion, Athens.

Tzartzanos, A. 1946-63. *Νεοελληνική σύνταξις (της κοινής δημοτικής)* [Modern Greek Syntax (of dimotiki)]. Organismos Ekdoseos Didaktikon Vivlion, Athens.

A New Greek Corpus

Dimitra Alexandridou[(1)], Anna Anastassiadis-Symeonidis[(2)]

[(1)]*LDI CNRS UMR 7187, University of Paris 13*, Paris, France.
dalexandridou@univ-paris13.fr
[(2)] *Aristotle University of Thessaloniki*, Thessaloniki, Greece
ansym@lit.auth.gr

Abstract

The objective of this paper is to present the development of a Greek corpus in a multilingual comparative context. Within the framework of natural language processing and information extraction, we examine its linguistic, informational and computational aspects. We develop a corpus building tool accordingly and we evaluate its results.

1 Introduction

Our broader objective is to develop an application for information extraction in Greek, French and English. We rely on linguistic analyses, formalized in electronic dictionaries and finite state automata. The application's performance depends, on one hand, on the suitability and the exhaustiveness of the analyses, on the other hand, on their formalization, that has to take into account the particularities of each language. We rely on very large corpora in order to carry out, evaluate and validate our analyses.

After examining the linguistic, informational and computational aspects of our context, as well as the existing corpora for the Greek language, we present the development of a dynamic web-based corpus.

1.1 Corpus characteristics

From a linguistic point of view, we are interested in the combinatory of lexical units on the phrase level and we seek for synchronic general language texts. The morpho-syntactic and syntactico-semantic analysis of the corpus will enable us to enrich our dictionaries and to detect potential neologisms. From an informational point of view, we seek for texts that are comparable in terms of register, topic and date of publication in order to carry out equivalent analyses in the three languages. From a computational point of view, texts need to be uniform in terms of encoding and structure.

1.2 Greek corpora

There are three electronic corpora of the Greek language: the Hellenic National Corpus (HNC)[1], the Corpus of Greek Texts (CGT)[2]and the corpus of the Portal for the Greek Language (PGL)[3]. They consist of 47, 30 and 4 million words respectively. Their content is summarized in the following tables:

[1] http://hnc.ilsp.gr/
[2] http://sek.edu.gr/
[3] http://www.greek-language.gr/

HNC Written texts	100%
Books	9,41%
Internet	0,32%
Newspapers	61,29%
Magazines	5,89%
Miscellaneous	23,08%

Table1. HNC Written Texts

CGT Oral texts	10%
Spontaneous discourse	1,67%
Public interviews	5%
Radio, TV discourse	3,33%

Table 2. CGT Oral Texts

CGT Written texts	90%
Literary books	16,67%
Informational books	16,67%
Academic discourse	16,67%
Internet Press : news	16,67%
Internet Press : opinions	16,67%
Official documents	6,67%

Table 3. CGT Written Texts

PGL Written texts	100%
Newspapers	100%

Table 4. PGLWritten Texts

2 Web as corpus

The web constitutes a very important collection of texts that can form the source of our corpus. It offers a vast diversity of languages, topics and registers; it is dynamic, therefore essentially synchronic; and, in terms of size, it constitutes the greatest available source of electronic texts.

Its inconvenience relies in retrieving and processing relevant information. Search engines are one of the most important means in web information retrieval, however, they condition queries and their results are open to an interpretation. Once the target page is retrieved, it needs to be preprocessed before it can be integrated in a corpus: information, such as the title, the date of publication or the text itself, is not uniformly marked across sites and needs a site- or even a page-specific instruction in order to be extracted (cf. infra **HTML**).

Nonetheless, online newspapers offer several advantages as compared to other sources on the web. From a linguistic point of view, each newspaper is uniform in terms of domain (general language), register (journalistic) and dialect (standard Greek). From an

informational point of view, its content is already categorized by topic allowing for a preselection of texts. From a computational point of view, all pages of a newspaper share a common information structure; once it is identified, information extraction can be systematized rendering online newspapers a perpetual[4]source of texts. Furthermore, the majority of newspapers index their content in a highly accessible format (cf. infra **RSS**) that is regularly updated.

Newspapers publish their content in two markup languages: HTML[5] and RSS[6]. Markup languages allow for the semantic annotation of a text and are syntactically distinct from the text itself. Semantic annotations can be predefined or not, depending on the markup language, and can contain presentation, procedural or descriptive information.

2.1 HTML

HTML documents use predefined annotations that provide all three kinds of information. The resulting structure however varies from page to page since annotations can be used in conjunction, and each block can be further subdivided according to the writer's needs. Thus, while it allows for a very accurate, personalized presentation of a document, it can complicate information exchange, and demands for a specific processing instruction in information extraction.

Systematizing text extraction from an HTML document consists in identifying the block containing the text in its integrity. In the case of online newspapers it is valid for all its articles. We provide two extraction methods, one that extracts a block without making any changes, that we call "keep", and one that allows choosing internal blocks that are not to be considered as part of the text (for example image captions or tables, see Figure 1), that we call "remove".

[4] That is, up to an eventual structure updating.
[5] http://www.w3.org/TR/html401/
[6] http://cyber.law.harvard.edu/rss/rss.html

Figure 1. HTML text blocks

Text blocks are usually further subdivided to include style, script or link annotations:

```
<!DOCTYPE HTML PUBLIC "-//W#C//DTD HTML 4.01">
<html>
 <head>...</head>
 <body>
  .
  .
  <div id="arttext">
   <h4><P>Παγωμένη Κιβωτός για δείγματα γενετικού υλικού ειδών που
έχουν εξαφανιστεί ή που βρίσκονται στα όρια της εξαφάνισης
λειτουργεί στο Πανεπιστήμιο του Νότιγχαμ της Αγγλίας, ούτως ώστε
στο μέλλον να μπορούν εξαφανισμένα είδη να αναβιώσουν μέσω
κλωνοποίησης για να διατηρηθεί η βιοποικιλότητα του πλανήτη Γη! </
P></h4>
   <div class="photo2">...</div>
   <P>Προς αυτή όμως την κατεύθυνση, της διατήρησης της
βιοποικιλότητας, κινούνται αργά και οι ηγέτες των χωρών, όπως
τουλάχιστον φαίνεται από τη νέα Συνθήκη για τη Βιοποικιλότητα, για
την οποία συμφώνησαν αντιπρόσωποι των κρατών-μελών του ΟΗΕ, σε
διάσκεψη που έγινε για το θέμα.</P>
   <P><STRONG>Δείγματα<BR></STRONG>Τα τελευταία δείγματα γενετικού
υλικού που έχουν φτάσει στην Παγωμένη Κιβωτό είναι της αρκούδας
της Μαλαισίας (Helarctos malayanus), η οποία έχει φτάσει πολύ κοντά
στην εξαφάνιση.</P>
   <P>Όπως τόνισε η καθηγήτρια Αν Κλαρκ, μια από τους ιδρυτές της
Παγωμένης Κιβωτού, το πρόγραμμα αυτό θα μπορούσε να χρησιμεύσει
σαν σχέδιο Β' όταν όλες οι άλλες προσπάθειες για τη διάσωση των
υπό εξαφάνιση ειδών έχουν αποτύχει. Συγκεκριμένα σε δηλώσεις της
ανέφερε: «Όπως με την Ντόλι, το κλωνοποιημένο πρόβατο, παρότι οι
επιστήμονες δεν θέλουν να μιλούν ανοιχτά γι' αυτό, υπάρχει ένα πολύ
πιθανό ενδεχόμενο στο εγγύς μέλλον να μπορούμε να φέρουμε πίσω στη
ζωή ή να σώσουμε από την εξαφάνιση ένα είδος». </P>
  </div>
  .
  .
 </body>
</html>
```

Figure 2. text annotations

Although this information mightprove useful for linguistic analysis, normalizing the text facilitates future processing. The remaining annotations are considered as paragraph markers.

2.2 RSS

Extracting information such as the title or the date of publication from an HTML page would require a similar process, i.e. retrieving the appropriate blocks for each newspaper. RSS feeds are designed to facilitate such information exchange. They provide a standardized format for the metadata of a newspaper and each of its articles. The title, the date of publication and the web address (URL) are obligatory blocks for each article, providing a uniform structure for all online newspapers. Usually a different feed is provided for each topic.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <rss version = "2.0">
3   <channel>
4    <title>v4.ethnos.gr - E-LIFE</title>
5    <link>http://www.ethnos.gr/rss.asp?
·  catid=11808&amp;subid=20120&amp;tag=8356</link>
6    <item>
7     <title>Παγωμένη Κιβωτός για είδη υπό εξαφάνιση</
·  title>
8     <link>http://www.ethnos.gr/article.asp?
·  catid=11900&amp;subid=2&amp;pubid=40404973</link>
9     <pubDate>Sat, 30 Oct 2010 09:57:31 GMT</date>
10   </item>
11  </channel>
12 </rss>
```

Figure 3. RSS feed

One of the most important advantages of RSS feeds is that they offer a complete index of the newspapers' current articles, facilitating updates. Since RSS feeds can be partially updated we verify that an article is not already processed using its URL as its unique identifier.

Indexing the articles of a newspaper that does not provide RSS feeds would require the implementation of a crawler in order to extract every link from every page and test whether it corresponds to an article or not, a far more resource demanding process.

2.3 Data normalization

The extracted data is normalized in a UTF-8 encoded XML document. XML[7] is a non-predefined markup language that allows for the specification of content appropriate structures and annotations. The structure adopted for the Greek corpus contains a building block for each article (*item*). Each principal block is further subdivided to include the name of the newspaper (*ejournal*), the article's web address (*guid*), its topic (*category*), its date of publication (*date*), its headline (*headline*), an eventual subhead, and the text (*text*) divided in paragraphs (*p*) (see Figure 4).

[7] http://www.w3.org/TR/xml/

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE configuration SYSTEM "corpus.dtd">
<corpus>
 <item>
  <category>E-LIFE</category>
  <date>Sat, 30 Oct 2010 09:57:31 GMT</date>
  <ejournal>TO ΕΘΝΟΣ</ejournal>
  <guid>http://www.ethnos.gr/article.asp?
catid=11900&amp;subid=2&amp;pubid=40404973</guid>
  <headline>Παγωμένη Κιβωτός για είδη υπό εξαφάνιση</headline>
   <text>
    <p>Παγωμένη Κιβωτός για δείγματα γενετικού υλικού ειδών που
έχουν εξαφανιστεί ή που βρίσκονται στα όρια της εξαφάνισης
λειτουργεί στο Πανεπιστήμιο του Νότιγχαμ της Αγγλίας, ούτως ώστε
στο μέλλον να μπορούν εξαφανισμένα είδη να αναβιώσουν μέσω
κλωνοποίησης για να διατηρηθεί η βιοποικιλότητα του πλανήτη Γη!</p>
    <p>Προς αυτή όμως την κατεύθυνση, της διατήρησης της
βιοποικιλότητας, κινούνται αργά και οι ηγέτες των χωρών, όπως
τουλάχιστον φαίνεται από τη νέα Συνθήκη για τη Βιοποικιλότητα, για
την οποία συμφώνησαν αντιπρόσωποι των κρατών-μελών του ΟΗΕ, σε
διάσκεψη που έγινε για το θέμα.</p>
    <p>Δείγματα</p>
    <p>Τα τελευταία δείγματα γενετικού υλικού που έχουν φτάσει στην
Παγωμένη Κιβωτό είναι της αρκούδας της Μαλαισίας (Helarctos
malayanus), η οποία έχει φτάσει πολύ κοντά στην εξαφάνιση.</p>
    <p>Οπως τόνισε η καθηγήτρια Αν Κλαρκ, μια από τους ιδρυτές της
Παγωμένης Κιβωτού, το πρόγραμμα αυτό θα μπορούσε να χρησιμεύσει
σαν σχέδιο Β' όταν όλες οι άλλες προσπάθειες για τη διάσωση των
υπό εξαφάνιση ειδών έχουν αποτύχει. Συγκεκριμένα σε δηλώσεις της
ανέφερε: «Οπως με την Ντόλι, το κλωνοποιημένο πρόβατο, παρότι οι
επιστήμονες δεν θέλουν να μιλούν ανοιχτά γι' αυτό, υπάρχει ένα πολύ
πιθανό ενδεχόμενο στο εγγύς μέλλον να μπορούμε να φέρουμε πίσω στη
ζωή ή να σώσουμε από την εξαφάνιση ένα είδος».</p>
   </text>
  </item>
</corpus>
```

Figure 4. XML structure

3 Corpus building tool

The manual tasks of the corpus building process consist in retrieving online newspapers and their corresponding RSS feeds and identifying the text block for each newspaper. This information serves as input for the corpus-building tool and is stored in an XML configuration file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE configuration SYSTEM "configuration.dtd">
<configuration>
 <item name = "TO ETHNOS" file = "to_ethnos.xml">
  <html subhead = "td class srcsubtitle" text = "div id
arttext" method = "remove">
   <tag>div id ban160</tag>
   <tag>div class photo2</tag>
  </html>
  <rss category = "E-LIFE" url = "http://www.ethnos.gr/
rss.asp?catid=11808&amp;subid=20120&amp;tag=8356"/>
 </item>
</configuration>
```

Figure 5. Tool configuration file

For each newspaper we include the web addresses of its RSS feeds, the annotation of the text block, and the eventual internal blocks that we may wish to remove. All data are normalized according to the XML standard.

Data processing is automated with a Perl script and consists in:

1. downloading each RSS feed
2. extracting information for each article
3. downloading each article (HTML)
4. extracting the text block
5. normalizing the data

Only valid and well-formed[8] documents are processed. The output is stored in a separate file for each newspaper.

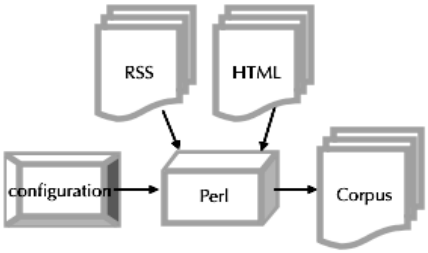


Figure 6. Corpus building process

4 The Greek Corpus

Seven online newspapers compose the Greek corpus: SentraGoal, Imerisia, Kathimerini, Ta Nea, To Vima, Cinema and To Ethnos. It currently contains 60 million words and 12 main topics: news, economy, politics, sports, technology, science, environment, traveling, culture, education, health and cooking. A detailed topics list is provided in Appendix A.

We automatically evaluated the encoding, the formalization and the validity of the corpus with a Perl script. We manually evaluated the tool's extraction efficiency in a sample of 100 texts for each newspaper: it extracts the text in its integrity and correctly annotates the paragraphs.

5 Perspectives

We plan to enrich the Greek corpus with additional newspapers and proceed to its morpho-syntactic and syntactico-semantic annotation. The French and the English corpus are under development.

[8] A document is considered well formed when it conforms to the general specifications of the markup language, valid when it conforms to its structure and annotation specifications.

References

Γούτσος, Δ., King, P.& Χατζηδάκη, Ρ. 1995. *«Η χρήση των corpus στη λεξικογραφία και περιγραφή της Νέας Ελληνικής».* In Μελέτες για την Ελληνική Γλώσσα. Πρακτικά της 15ης ετήσιαςσυνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης, 11-14 Μαΐου 1994. Θεσσαλονίκη: Αφοί Κυριακίδη, 843-854.

Γούτσος, Δ. 2003. *«Σώμα Ελληνικών Κειμένων: Σχεδιασμός και υλοποίηση».* In Πρακτικά του 6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας, Πανεπιστήμιο Κρήτης, 18-21 Σεπτεμβρίου 2003.

Amitay, E. 1999. *« Anchors in Context: A corpus analysis of web pages authoring conventions »*, In Words on the Web - Computer Mediated Communication. Lynn Pemberton and Simon Shorville: Intellect Books, UK.

Fairon, C., Sincler, J.V. 2006. « *I'm like, 'Hey, it works!': Using GlossaNet to find attestations of the quotative (be) like in English- language newspapers* », in A. Renouf and A. Kehoe (eds). The Changing Face of Corpus Linguistics. Language and Computers 55. Rodopi, Amsterdam/New York, NY, pp. 325-336.

Benoît, H., Nazarenko, A. & Salem, A.1997. *Les linguistiques de corpus*. Paris : Armand Colin & Masson.

Issac, F., Fouquere, C. 2003. « *Corpus issus du web : constitution et analyse informationnelle* ». Revue Québécoise de Linguistique 32 (1) , pp. 111-134.

Issac, F., Hamon, T., Fouquere, C., Bouchard, L., Emirkanian, L. 2001. *"Extraction informatique de données sur le web"*. DistanceS 5 (2) , pp. 195-209.

Gulli, A., Signorini, A. 2005. « *The indexable web is more than 11.5 billion pages* » in International World Wide Web Conference, pp. 902-903.

Pincemin, B., Issac, F., Chodkiewicz, C. 2006. « *The XML/TEI Human Rights Corpus* ». TEI Annual Members' Meeting, Vancouver October 27 and 28.

Resnik, P. 1998. *Parallel strands: A preliminary investigation into mining the web for bilingual text*. In Proceedings of 1998 Conference of the Association for Machine Translation in the Americas.

Wooldridge, R. 2004, *« Le Web comme corpus d'usages linguistiques »*, in Cahiers de lexicologie, 85: 209-25.

Appendix A. Detailed topics list

SENTRAGOAL
Football
Basketball
Volleyball
Other sports
Athletics
Handball
Water sports
TO ETHNOS
News
Politics

Society
World
Economy
Opinions
Columnists
E-life
Science
Health and life
Cooking
Environment
Technology
Outdoor
Outdoor-fishing
Arts
Civilization
Cinema
TV
Showbiz
Book
Work
Education
Traveling
Holidays
Cinema
News
Movies
Faces
In theatres
Box office
DVD
TO VIMA
Politics
Economy
Society
World
Opinions
Civilization
Science
Sports
Books
IMERISIA
News
Economy
Companies
Business
Stock markets
Greece
World
Real estate
Opinions
Interviews
Market indicators
Sectors

Packets	
Interest rates	
Currency	
Goods	
Environment	
Technology	
Digital life	
Telecoms	
Gadgets	
Science	
Lifestyle	
Goodlife	
Culture	
Columns	
Luxury	
KATHIMERINI	
News	
Economy	
Business	
Politics	
Greece	
World	
Civilization	
Sports	
Technology	
Science	
Home	
Passport (paper version)	
TA NEA	
News	First page
Greece	World
Economy	Civilization
Sports	Opinions
Icons	Car news
Holidays	Books
Legal news	Employment