NooJ 2010

Komotini - Greece 27-29 May

ABSTRACTS

Department of Greek Philology
Democritus University of Thrace

Contents

Automatic Transformational Analysis with NooJ

Max Silberztein, Université de Franche-Comté

max.silberztein@gmail.com

We will present NooJ's new functionalities that perform the automatic generation of paraphrases, from a given text and a corresponding syntactic grammar. As opposed to previous devices, such as INTEX's or NooJ's finite-state transducers (FSTs) and recursive transition networks (RTNs), NooJ's new operator does not force users to design only one-to-one transformations, and the description of a transformation is no longer oriented: the same grammar that displays the correspondence between an active and a passive sentences can be used to produce both an active sentence from a passive one, and a passive one from an active one. Moreover, describing a transformation is straightforward, does not require a new formalism or new devices, as it can be achieved using ordinary syntactic grammars.

A French-English MT system for Computer Science Compound Words

Farida AOUGHLIS, Université Mouloud Mammeri,
Tizi Ouzou, Algérie
fariyamo@yahoo.fr

Automatic translating software needs increasingly significant and varied terminological resources. They can be simple lists of terms that are more or less structured (structured indices, thesaurus, lexical networks) used by automatic indexing systems or for information retrieval or more documented terminological reference frames.

Today, industrial and large-audience Machine Translation software are still producing poor quality results.

For the technical languages or of speciality, work remains to be made to build electronic dictionaries for NOOJ and MT systems for terminology.

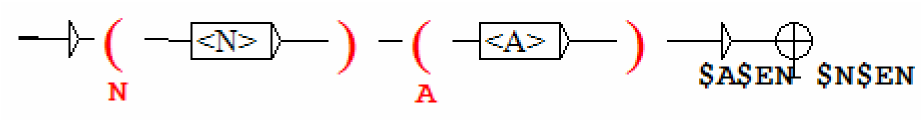
Our aim is to develop a MT system for computer science compound words, from French to English. We have a computer science dictionary for French compounds, a computer science dictionary for English compounds.

- Our computer science compound words French dictionary INFO_COMP contains actually more than 10 000 compounds words, 30 000 terms are collected and will be added to the dictionary.
  We extract a part of the dictionary for the MT test and add the translation in English for each entry.

- Actually our computer science compound words English dictionary ENG_COMP contains some entries, with a French translation.

- We are studying the possible translations for the compounds, we see that more translation are not 'word to word' translations for example:

  Mémoire vive has 4 translations => RAM, Random access memory; Computing store; Random access storage; Read write memory
  Programme de chargement => Boot; Load program; Loader
  Mémoire auxiliaire => Auxiliary memory , here AN => NA in English
- We build some syntactical translation grammars with for input Fr and output En, we give the translation grammar for *mémoire auxiliaire*:



- When we finish this work, we hope orient us to a MT system for sentences with compounds.

Key words: MT, Terminology, Electronic dictionary, Compound Words.

Extraction of relations between Arabic Named Entities using NooJ platform: Case of sport domain

Hela Fehri (1), Odile PITON (2) and Abdelmajid BEN HAMADOU (3)
(1), (3)MIRACL, Sfax university, Tunisia
hela.fehri @ fss.rnu.tn, Abdelmajid.benhamadou @ isimsf.rnu.tn
(2) SAMM, Sorbonne university, France
Odile.Piton@univ-paris1.fr

The objective of this research is to develop a system for the extraction of semantic relations between Named Entities concerning the sport domain. In fact, extracting relations between Named Entities in a text is a very important applicative research domain. It can be, for example, integrated in a question/answering system in order to retrieve the appropriate information requested by the users.

Relations between named entities can be binary (eg: **Located_in** (place: SPORT_LOC, location: TOP), this relation can be instantiated as follows Located_in (درعا ,ملعب تشرين) wish means that the place "*maalab techrin*" is located in the town "*Deraa*"), or n-ary (i.e., evenement) eg: **Competition** (team1: PN, team2: PN, date: DAT, stadium: LOC, location: TOP)) wish can be instantiated as (الترجي التونسي *el_taraji el_tounsi*, النادي الصفاقسي *el_nadi el_sfaxi*, ملعب الطيب المهيري , 2009 *malaab el_taeib el_mhiri*, صفاقس *sfax*)). Arguments with capital letters represent named entity categories.

The extraction is performed on two steps: Identification of the named entities and Detection of relations between identified named entities.

An implementation of the recognition and detection of semantic relations process is made using the platform NooJ. Already, we have developed tools for recognition of named entities like Name of Sport Locations (SPORT_LOC), Toponyms (TOP), Dates (DAT), name of proper names (PN). The identification of the Located_in relation is performed and the rest of semantic relations between named entities are in progress. So, to reach our objectives, some transducers should be added.

The present work is done between the MIRACL laboratory Sfax University and SAMM laboratory (Sorbonne University).

Keywords: Named entity, semantic relations, sport domain, recognition.

Vers le traitement du judéo-espagnol par NooJ

Ana Stulic-Etchevers(1), Soufiane Rouissi (1) Duško Vitas(2), Cvetana Krstev (2)
(1)AMERIBER, CEMIC-GRESIC, Université Michel de Montaigne, Bordeaux 3
ana.stulic@u-bordeaux3.fr, soufiane.rouissi@u-bordeaux3.fr
(2), Université de Belgrade
vitas@matf.bg.ac.rs, cvetana@matf.bg.ac.rs

Dans ce travail, nous examinons la possibilité de l'usage et de l'adaptation des ressources développées dans le système NooJ pour le traitement de la langue espagnole des textes judéo-espagnols dans le cadre du projet *Corpus numérique judéo-espagnol*, mené par le groupe recherche AMERIBER de l'Université Michel de Montaigne-Bordeaux 3 en collaboration avec l'Université de Belgrade. L'objectif du projet est de rendre disponible à des fins de recherche les textes en judéo-espagnol, langue parlée par les juifs séfarades expulsés d'Espagne en 1492. Les documents séfarades se caractérisent par une très grande diversité en ce qui concerne leur contenu, situation communicative et pays dans lesquels ils ont été produits, mais aussi pour ce qui est du système d'écriture employé : les documents judéo-espagnols qui constituent notre corpus ont été rédigés en caractères hébreux, latins et cyrilliques. Les documents traditionnels judéo-espagnols sont rédigés en une adaptation d'écriture hébraïque où les voyelles ne sont que partiellement indiquées, c-à-d toutes les voyelles sont notées, mais un seul graphème, yod, est employé pour noter deux voyelles palatale /e/ et /i/; et un seul graphème, waw, est utilisé pour les voyelles vélaires /o/ et /u/. Si pour un lecteur averti la lecture d'un texte judéo-espagnol en écriture hébraïque est relativement aisé, le texte numérique judéo-espagnol en écriture hébraïque ne permet pas l'usage immédiat des resources développées pour l'espagnol. Nous proposons, par conséquent, un procédé qui permet d'obtenir une version normalisée du texte où les voyelles sont interprétées à l'aide d'un dictionnaire développé spécifiquement pour le judéo-espagno.

Mots-clés : judéo-espagnol, NooJ

Processing Greek frozen expressions with Nooj

Michalis Anastasiadis[1], Lena Papadopoulou[2], Zoe Gavriilidou[1],
1 zoegab@otenet.gr Democritus University of Thrace
2 lepapad@hotmail.com Universidad Autonoma de Barcelona

The automatic recognition, representation, processing and translation of frozen expressions remain an obstacle in NLP research. This is due to their form fixedness, idiosyncratic meaning and sometimes discontinuity of their elements. In Nooj "simple words and multi-word units are processed in a unified way: they are stored in the same dictionaries, their inflectional and derivational morphology is formalized with the same tools and their annotations are undistinguishable from those of simple words." (Silberstein 2007)

This paper deals with the automatic processing of Greek frozen expressions. Fist, we attempt a classification of the entries. Then we describe the morpho-syntactic properties of our data. Finally we demonstrate the format of the Greek frozen expressions' dictionary as well as the graphs created for their processing and automatic translation in French. Our work is based on a corpus of 5000 entries whose compilation took under consideration previous work of Fotopoulou (1993) and Moustaki (1995).

References:

Fotopoulou, A. (1993), *Une classification des phrases à complements figés en Grec Moderne,* thèse de Doctorat, Université Paris VII .

Moustaki, A. (1995), Les expressions figées être Prep C W en grec moderne, thèse de Doctorat, Université Paris VII .

Silberstein, M. (2007), "Complex annotations with Nooj", *Proceedings of the 2007 International Nooj Conference*, Cambridge Scholar Publishing, pp. 214-227.

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) (3) MIRACL-University of Sfax, Tunisia
kais.haddar @ fss.rnu.tn, Abdelmajid.benhamadou @ isimsf.rnu.tn

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 fosters 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.

The choice of the representation of the formal framework is very important especially for the generation phase. Thus, the formal framework can improve the translation of NEs. Notice that its feeding (that means the assignment of values to features) can be performed using the recognition process to move NEs to a transfer approach in order to facilitate the reuse and multilingualism.

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 generalize the process of recognition and translation of NEs, whatever the domain and the chosen type of the domain hierarchy are.

The representation framework that we propose is based on the structure of features independently of lexical categories. Each structure represents an NE. However, the structure can be nested and complex as NE may contain one or more NEs. The features of the model are NEs (simple or compound), trigger words, contexts and ends of NEs and NE elements. As noted earlier, this framework will be exploited in the translation of NE. Indeed, the recognition phase assigns each feature a value that will be translated later.

The originality of our work lies in the fact that our representation model is very simple to implement in NooJ. Indeed, each NE will be transformed into a simple dictionary and each NE will be structured into a grammar. The NEs which are structured and nested represent subtransducers, while the other features represent transducer nodes. Therefore, we can identify all the dictionaries and grammars needed to implement the process of recognition and translation.

Note also that all the existing work of translation using the platform NooJ adopt a semi-direct approach. However, the NE representation will allow us to move from the semi-direct translation to transfer translation. Indeed, we should take into account the separation of the adjusting step from translating word to word. This helps the promotion to the reuse of the needed grammars. Thus, the reuse of the resources becomes possible. It is sufficient to change inputs (e.g., dictionaries, morphological grammars) of syntactic grammars for the desired results.

Keywords: formal representation, translation, NE, transfer method.

Arabic POS tagging based on NooJ grammars and the Arabic morphological analyzer MORPH2

Nouha Chaâben Kammoun (1), Lamia Hadrich Belguith (1), Slim MESFAR (2),
(1) FSEGS, Univesity of Sfax, Tunisia
chb_nouha@yahoo.fr, l.belguith@fsegs.rnu.tn
(2) RIADI, University of Manouba, Tunisia
mesfarslim@yahoo.fr

Part-Of-Speech tagging (POS tagging) is the process by which a specific tag is assigned to each word of a sentence to indicate its function in the specific context [Jurafsky & Martin, 2008]. It is an important task for any natural language application. In the sate of the art, we distinguish three POS tagging approaches [Mohamed Elhadj, 2009]: the rule-based approach, the statistical approach and the hybrid approach. Rule-based approach uses a knowledge base of linguistic rules to assign tags to words. Statistical approach uses machine learning techniques to process POS tagging. Hybrid approach uses a combination of the two previous approaches to assign tags to words.

In this paper, we propose a rule-based POS tagging task based on NooJ grammars and the

Arabic morphological analyzer MORPH2 ([Belguith & Chaâben, 2006]; [Chaâben & al., 2010]). In a first step, the Arabic text is morphologically analyzed with MORPH2. As input, the system accepts an Arabic word, a sentence or a text. This input can be non-vocalized, partially vocalized or fully vocalized. As output, an XML file is generated. This file contains all information extracted by MORPH2 (i.e. all possible morphological features for each word). This XML file is, then, used as input to NooJ platform on which NooJ grammars are applied to filter morphological solutions provided by MORPH2. As a result, each word in the text has a single morphological solution. To be used in other natural language processing levels, the result is exported from NooJ platform to an XML file.

Keywords: NLP, Arabic language, POS tagging, morphological analysis.

References:

Belguith Hadrich L. and Chaâben N., 2006. Analyse et désambiguïsation morphologiques des textes arabes non voyellés. Actes de la 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006), pp. 493-501.

Chaâben Kammoun N., Belguith Hadrich L. et Ben Hamadou A., 2010. The MORPH2 new version: A robust morphological analyzer for Arabic texts, to appear in JADT'2010.

Jurafsky D. and Martin J.H., 2008. Speech and Language Processing: An Introduction to

Speech Recognition, Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, ISBN: 10: 0131873210, pp: 1024.
Mohamed Elhadj Y.O., 2009. Statistical Part-of-Speech Tagger for Traditional Arabic Texts, Journal of Computer Science 5 (11): 794-800, ISSN 1549-3636.

Colors in Catalan and their treatment with NooJ

Puig Portella Marcel
Autonomous University of Barcelona, Marcel.Puig@uab.cat

FLexSem Group is a research group in linguistics of the French and Romanic Philology Department in the Autonomous University of Barcelona. Main fields of research embrace semantics, lexicology, and phonetics for speech pathology

In the framework of the group, we are working in semantics and combinatory for sentences incorporating color semantically related lemmas in Catalan language.

Color terminology is especially ambiguous in a text sequence, mainly since so many meanings are related to it. Color adjectives are firstly and very often derived from other semantic entities that still preserve the same root terminology, as for instance in *orange* or *violet*. On the other hand, the same terminology is very often used in naming entities only partially associated with the semantic of colors, as for instance in (egg) *white*, or *yellow*, *green* and *red* for traffic lights, in a sort of metonymy. Eventually, to the traditional terminology of colors, other, perhaps less considered but not less common, terms have been incorporated, which steadily retain the main semantics of their originary fields, like in *coffee* or *milk* colors. Only context consideration will be able to solve those challenges.

Using Nooj as an exploring tool, we intend to find out which structures can correlate entirely and properly to the domain of colors, segregating them from other possible semantic values. In productivity, a syntactic grammar will be of a great use, filtering the non suitable lemmas, avoiding ambiguity and therefore allowing a more accurate translation.

A disambiguation grammar is intended to be associated with the Catalan dictionary (Sastre, 2007). A hypercategory ***Color*** will be created, including any single lemma possibly related to the general semantics of colors. Hence the grammar will be able to recognise the real color meanings out of the general and less restricted color terminology. This microlexicon microstructure will be the first step in the disambiguation of the color related sentences, being the grammar the second one. The grammar will be working incorporating the hyperclass of colors.

As for the grammar itself, we are currently working with different text sequences and collocations, considering verbal, adjectival and nominal structures, with their correspondent inflectional forms. Some of the co-occurrences in the colors' domain are very simple, though very productive, as for instance:

Taronja clar, En= light orange
Taronja fosc, En= deep orange vs. Taronja, En= orange
Others, related to nominal or verbal constructions:
<Pintar> de <Color> (i <Color>) → <to paint> <Color> (and <Color>)
Color <brillar> → Color <to shine>
<Taca> <Color> (i <Color>) → <Color> [(and) <Color>] <spot>

Using Nooj for corpora incorporation as well as in context disambiguisation we intend to create a more precise approach to the semantics of colors, perhaps discriminating other semantic fields worth working on in a near future.

Keywords: lexicography, disambiguation, colors.

A comparative study of table-tennis vocabulary in Greek and English

Nikos Siklafidis[1], Papadopoulou Eleni[2]
(1) Democritus University of Thrace, ns9292@helit.duth.gr
(2) Autonomous University of Barcelona, lepapad@hotmail.com

The aim of the present study was to enrich the main lexical resources of the Greek Nooj Module (Gavriilidou, Papadopoulou and Chatzipapa 2008) with the construction of two bilingual (Greek-English) dictionaries (one for simple and one for compound nouns) including table-tennis terms, extracted from a specialized table-tennis corpus. Our work was based on previous research based on the lexicon grammar theory concerning the vocabulary of tennis in Greek, French and English (Sklavounou 1993) and the vocabulary of football in Greek and French (Moustaki and Dimitriadi 2006).

The theoretical framework that underlies our work is the *Lexicon-Grammar* (Gross, 1975), the *Classes of objects* (Gross, 1992) and the *monolingual coordinated dictionaries* (Blanco, 2001).

First, we will describe the microstructure of our dictionaries, which contains the grammatical category, the morphological category, the morphosyntactic feature, the class of objects, the domain and the equivalent translation of each lemma in English:

ρακέτα,N+FLX=N24+Conc+Equip+Ttennis+EN=paddle
ψείρα κοντά στο φιλέ,N+FLX=N24INININ+Abst+Tecnique+Ttennis+EN= drop shot

Then we will present a series of local grammars that can recognize various structures, such as Arg0VArg1, which can be analysed, into Det+<player>+V+Det+A<technique> (*ο πιγκπονίστας επέστρεψε ένα γρήγορο σερβίς με τόπσπιν*: *the table-tennis player returned a fast topspin service*).

Keywords: Greek and English language, specialized dictionary, table-tennis.

References:

Blanco. X. 2001. *Dictionnaires électroniques et traduction automatique espagnol français*. Langages 143. Larousse. Paris.

Gavriilidou. Z. Papadopoulou. E. Chatzipapa. E. 2008. *The New Greek NooJ Module: morphosemantic issues*. in Blanco, X.; Silberztein, M. (eds). *Proceedings of the 2007 NooJ International Conference*. Cambridge. Cambridge Scholars Publishing. p. 96-102.

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

Gross. M. 1975. Méthodes en syntaxe. Paris. Hermann.

Moustaki. A. Dimitriadi. A. 2006. Un lexique-grammaire du football. Étude contrastive du grec modern et du français. Linguisticae Investigationes. Amsterdam John Benjamins.

Sklavounou. E. 1993. *Un Lexique-Grammaire trilingue de noms composés (grec-français-anglais)-Application au vocabulaire spécialisé du tennis*. D.E.A.. 2 Vol. PARIS. Université Paris 8.

Event extraction in business news using NooJ

Božo Bekavac (1), Kristina Vučković (2), Željko Agić (2), Marko Tadić (1)
University of Zagreb, Faculty of Humanities and Social Sciences,
(1) Department of Linguistics (2) Department of Information Sciences
bbekavac@ffzg.hr, kvuckovi@ffzg.hr, zagic@ffzg.hr, mtadic@ffzg.hr

In this work we use NooJ as a core part of a broader system designed for Croatian business news analysis. The system consists of three parts: web crawler, event analyzer and event presentation module.

Documents are retrieved from ten different web sites (using different encodings) and filtered according to the user defined keywords. Typical keywords are names of 10-30 companies in which the user is interested. Filtered documents are stored locally and processed with NooJ. The core of the system uses NooJ grammars aiming to recognize 3 different predefined events: a) company's A takeover of company B, b) sharp rise or fall of certain company stock, c) certain company launching a promising product on the market.

NooJ local grammars are designed to recognize relationships between companies, stocks and products described in business news. Output of processing is recognition of subjects involved in certain event and type of predefined event mentioned above. Detected subjects and type of relation are further encoded as transducer's output tags and sent to event presentation module.

The results of processed texts will be evaluated through precision, recall and f-measure and analysis of most common mistakes will be discussed.

Keywords: event extraction, event recognition, NERC, Croatian language, NooJ

Terminology extraction and processing in Arabic language with NooJ system

Mohammed Lahalal Lahlal_benslimane@yahoo.fr
Azeddine Rhazi ourzagh@gmail.com

The general objective of this study is to clear up the relative importance of extracting Arabic language terms in the identification of composed of simple and multi-word from a specialized Arabic corpora. We use locale grammar implemented in INTEX linguistic tool (Max Silberztein 1993) to extract terms, the classification begins by the clustering technique, trough similar to hierarchical ascending classification based on subset of term units (Ibekwe_SanJwan 1998a & 1999).

The problem of extraction and terminological treatment is a lexicological phenomenon that has required an important work in NLP, which began with the automation of terms after their manual classification. We want to improve NooJ system parsing by means of statistical tools to investigate related tasks such as tagging, morphological analysis of the term that is the main challenge of Arabic language text lexical analysis.

In this paper, we'll take steps to give an overview on the effectiveness of systems based on knowledge (IKarus system), the notional multilingual network and demonstrate how the extraction of simple and compound words in Arabic, which really was not before, using the software NooJ.

Extraction of sequences that may contain terms is done by application of finite states transducers that allows us to extract the noun phrases, from the more complex to the atomic sequence from a given corpus, the specialist field can therefore manually validate certain candidate terms, that is, subsequently, an old method considered as automatic classification.

Our goal is the performance of NooJ system to extract and analyze the sequence of the corpora studied in Arabic.

An important problem faced by the NooJ system in describing natural languages on a wide scale is that of the finite between multi-word units (that must be described in dictionaries) and free word sequences (that must be described by syntactic grammars) it is evident for those problem automatic analysis of Arabic texts, that is the following examples : *safha bayda* (clean slate: freedom from commitments) we cant say *safha* without *bayda*.

To annotate noun phrases head of this term you must take contexts into account in order to disambiguate and filtering out the annotation mentioned in the text. By the way Arabic marked by the inflectional morphology witch demand a derivational one.

We develop a tool for treatment of terms by evaluating the system.. this tool makes it possible to identify the term from the corpora measuring the length of the term, to segment a heterogeneous text in homogeneous category by coding the sequence (noun, phraseology, idioms ..) we can use this tool in several different applications of NooJ environment. The tool is as an algorithm implemented by finite states transducers to generate and built an Arabic interactive dictionary of terms.

We propose a description of the environment NooJ. It will only matter of terminology extraction step.

Keywords: NooJ system, finite states transducers, NLP, Arabic language, simple and complex terms, finite states transducers.

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
[2]Swedish School of Library and Information Science, University College
Boras/Göteborg University

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, such as actant theory (i.e., actants are behavioral patterns in a story situation; one and the same actor can serve as different kinds of actant depending on the situation), or Propp's character functions[1] (i.e., there exists a limited set of prototypical sequence of 'character functions' such as "Conflict", "Kidnapping", "Test of Hero", etc.). Several computational models have targeted the processing of fairy tales, however, what is currently missing from such resources is specification about which actual linguistic elements are to be associated with a model's constituents (e.g. concepts, relations).

We attempt to process folk tale narratives to obtain different types of invariants. The importance of a text analytical approach based on language technology is to regard actors (roles) as slots filled in by particular names as values, which leaves one room to generalize the plot to a metalevel where, in turn, actions will be slots filled in by their values as well; for example, to extract the string '<Donor> sheltered <Hero> under its banks of pudding', where lexical items corresponding to character roles ("the river", respectively "her") are masked. The goal of such an approach is to find minimal actor-action-actor combinations whose particular value configurations result in specific subtypes of a tale.

2 Processing Fairy tales

[1] Propp, V.J. (1968) Morphology of the folktale. University of Texas Press: Austin.

We work on tales from the Afanasiev collection[1]; these are English translations of the Russian originals. Below is an excerpt from "Nikita the Tanner".

```
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. (...)
```

Characters provide a structure for the plot; to mention but a few:

- Hero: a character that seeks something;
- Villain: who opposes or actively blocks the hero's quest;
- Donor: who provides an object with magical properties;
- Princess: acts as the reward for the hero and the object of the villain's plots;
- Her Father: who acts to reward the hero for his effort.

2.1 Semantic Mapping to Character Roles

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; e.g. "`a 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 the PRINCESS to go to the VILLAIN .`" By manually performing a pilot character role assignment procedure, we identified a number of difficult or controversial semantic representation and mapping issues.

- Role transitions take place during the course of a story, e.g. the imp (a sort of demon in Nikita's story) sometimes acts as a Helper, at other times as Donor, and could at certain points also be labelled as False Hero.
- Some tales feature more than one instantiation of one and the same role, e.g. both the "Swan Geese" and "Baba Yaga" are Villains.
- Multiple persons, such as "tsar and tsarina" map to the role of Father (i.e., the assignment simplifies the actual multiple person involvement), whereas a flock of birds (the swan-geese) map to Villain.
- Synonyms that stand for the same role occur in various distributions in a tale. For example, "the girl", "daughter", "maiden", "she" all map to Hero, whereas "she" in certain passages has to be masked as Villain, because it refers to the witch.

2.2 Employing NooJ for Identifying Character Entities and Actions

In further work, we plan to address these issues by experimenting with the possibilities NooJ can offer. The ultimate goal is to identify the core (e.g.: "Donor sheltered the Hero"; "the Father rewarded the Hero", etc.) and periphery of lexical

[1] Afanas'ev, A. (1945). Russian fairy tales. Pantheon Books: New York.

units that can be associated with a certain Proppian function. We plan to initiate a module in NooJ that holds the lexicon of folk tale characters and actions, by processing English and Hungarian data.
In a subsequent step, we investigate the utility of basic linguistic information, like lemma, part of speech, or phrase boundaries, for the characterization of such subsentential elements by local grammars. Generalizing the results will shed light on the computational applicability of Propp's method for detecting genre-specific narrative units based on corpus evidence.

Version 4 Greek NooJ Module

Papadopoulou Eleni[1], Chatzipapa Elina[2], Karagiannopoulou Kiki[3],
(1) Autonomous University of Barcelona, lepapad@hotmail.com
(2) Democritus University of Thrace, elinaxp@hotmail.com
(3) Democritus University of Thrace,
karagiannopoulouk@gmail.com

Linguistic analyses of various Greek 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 text annotation of the Greek NooJ Module (Gavriilidou, Papadopoulou and Chatzipapa 2008), it was decided the construction of: a) two adverb dictionaries (one for simple and one for compound forms), b) a dictionary of acronyms and c) a dictionary of lexical units written with Latin characters that occur with high frequency in Greek texts. The features of each dictionary are described in the present paper.

For the construction of the two adverb dictionaries we were based on Catalá (2003). The entries of the simple adverb dictionary were further used for the construction of disambiguation grammars, such as for the disambiguation between adverbs and adjective forms ending in *–α* (i.e., *σύντομα*: *briefly* or *brief*). On the other hand, the entries of the compound adverb dictionary (e.g. *μια για πάντα* ‘once and for all’) resolved problems caused by frozenness and non compositionality.

For the elaboration of the acronym dictionary, we took under consideration previous work on Greek acronyms done by Vazou and Xydopoulos (2007).

Finally, the dictionary of lexical units written with Latin characters helped assign a grammatical gender and semantic information in this kind of lemmas.

Keywords: Greek language, lexicography, disambiguation, acronyms, adverbs.

References:

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

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

Gavriilidou. Z. Papadopoulou. E. Chatzipapa. E. 2008. *The New Greek NooJ Module: morphosemantic issues*. in Blanco, X.; Silberztein, M. (eds). *Proceedings of the 2007 NooJ International Conference*. Cambridge. Cambridge Scholars Publishing. p. 96-102.

Greek Professional nouns processed with Nooj

Chatzipapa Elina[1], Papadopoulou Eleni[2], Georgia Samani [3]

(1) Democritus University of Thrace, elinaxp@hotmail.com
(2) Autonomous University of Barcelona, lepapad@hotmail.com
(3) Democritus University of Thrace, georgia.samani33@gmail.com

This paper intends to present the study of the Greek professional nouns applied to the NooJ environment. The ability of Nooj, on the one hand, to maintain large coverage dictionaries by extending the macrostructure of an existent dictionary and, on the other hand, to introduce morphological grammars for the automatic recognition of the lemmas, constituted the base of the present study.

Our work started with the data collection of the Greek professional nouns. The application of the theoretical frame of *classes of objects* (G. Gross) contributed to the construction of the *Greek Professional Nouns Lexicon (G.P.N.L.)*. The selected appropriate operators for the professional nouns were: the verb "*επαγγέλλομαι*" (EN: - occupy) and the phrase "*ασκώ το επάγγελμα του*" (EN: - practice a profession). Since the lemma collected, the image of the microstructure of the G.P.N.L completed with the inflectional codification, the annotation of their syntactic-semantic feature (*Hum*), the class of object (<*Profession*>) and the translation equivalent in English.

In addition, the macrostructure consists of approximately 3.700 professional simple (χειρούργος, EN: - surgeon) and compound nouns (χειρούργος οδοντίατρος, EN: - dental surgeon). Some of them are historical professions, others are diatopics and some of them exist only in one of the two genders. For the automatic recognition of the professional nouns, a morphological grammar was constructed, based on a variety of suffixes -*της, -ευτής*, etc. confixes such as *–λόγος* or productive lexical roots forming compound such as *–παραγωγός* (EN: - *producer*), e.g *ελαιοπαραγωγός*//*olive-producer*.

Keywords: Class of objects <Profession>, Grammars, *G.P.N.L.*.

References:

Blanco, X. & Lajmi, D., (2004), Dictionnaire électronique français-espagnol-catalan-arabe des noms des profession et des métiers *Actes des Premières Journées*

Scientifiques des Réseaux de Chercheurs de l'Agence Universitaire de la Francophonie, Association d'Universités Francophones.

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.

Les mots de Mary Astell, étude lexicologique, grammaticale et sémantique d'un texte du XVIIe siècle au moyen de la plateforme linguistique NooJ

Helene Pignot
Odile Piton

The linguistic platform NooJ has great functionalities for lexicographic work. From a morphological viewpoint, thanks to the NooJ dictionary, it is possible to spot all the occurrences of words whose suffixation and prefixation were different in seventeenth-century English as those words are immediately recognized and listed as "unknown". Another (and more systematic) way to find these words is to apply certain commands, e.g. <A+MP="ous$">, to find all the occurrences of adjectives ending in “ous”.

From a semantic viewpoint, NooJ can help us identify the main themes of a text through the study of lexis. NooJ indicates the frequency in the use of all tokens and their co-occurrence with certain words. In Astell’s text the most commonly used words are GOD (always capitalized) and world (incidentally which may point to a tension in the text between two conflicting demands for women in the 17th century: find a place in the world and fulfill their spiritual destiny as God’s creatures, endowed with reason just as males). My colleague Odile Piton and I have been building up NooJ tools to process seventeenth century English (for example syntactic graphs wherever possible); now we would like to create a NooJ dictionary where the lexicon specific to the 17th century would be listed. This tool could be extremely useful for non-anglophone students and English students alike, who could read in their original version these beautiful but often difficult texts published in the seventeenth century, without being put off or impeded by their lexicon and their stylistic singularities.

Improved Parser for Simple Croatian Sentences

Kristina Vuckovic (1), Božo Bekavac (2), Zdravko Dovedan (1)
University of Zagreb,Faculty of Humanities and Social Sciences,
(1) Department of Information Sciences, (2) Department of Linguistics
kvuckovi@ffzg.hr, bbekavac@ffzg.hr, zdovedan@ffzg.hr

In this paper, authors will present the work that has been done to improve the existing syntactic parser. This work is a continuation of the work 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 enlight 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.

Keywords: parse trees, partial parser, simple sentences, Croatian language, NooJ, coordination, nominal predicate.

Learning the Greek Language via Greeklish

Anastasia D. Georgiadou (1), Alex S. Karakos(2), John Papaioannou(3)
(1) Democritus University of Thrace, anageorg@ee.duth.gr
(2) Democritus University of Thrace, karakos@ee.duth.gr
(3) Democritus University of Thrace, john.papaioannou@gmail.com

Learning Greek as a second (L2) or foreign (FL) language has been a pole of attraction for 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 a software for learning the Greek Language via Greeklish. Since, the basic vocabulary of a language is the beginning 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 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 that the user typed, and matches 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, a standalone program that accomplishes the transliteration of a whole text written in Greek to Greeklish and vice versa.

The aim is to provide a useful tool, a stand-alone 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.

Keywords: Greeklish, transliteration, dictionary, second language, foreign language, Greek

Deriving Nouns from Numerals

Sara Librenjak, Kristina Vuckovic, Zdravko Dovedan
University of Zagreb, Faculty of Humanities and Social Sciences,
Department of Information Sciences
sara.librenjak@gmail.com, kvuckovi@ffzg.hr, zdovedan@ffzg.hr

The paper discusses formation of numeral nouns 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.

Numeral nouns in Croatian language can be considered as a specific group of words since they possess a limited number of semantic roots (numeral part) and their endings are from the finite number of suffixes. Syntactically, they are used mostly in the same manner as nouns.

We believe that detailed description of numeral nouns will increase both precision and recall in recognition of noun phrases <NP> and prepositional phrases <PP>. This will be demonstrated by reapplying the local grammars for recognition of <NP> and <PP> chunks to the corpus.

Keywords: numeral nouns, productive morphology, noun phrases, prepositional phrases, Croatian language, NooJ.

Morphology based recognition of Greek verbs with Nooj

Angeliki Efthymiou, Democritus University of Thrace
Zoe Gavriilidou, Democritus University of Thrace
aefthym@eled.duth.gr
zoegab@otenet.gr

The Modern Greek suffixes *-ίζω, -(ι)άζω, -εύω, -ώνω, -άρω, -αίνω, -ύνω*, and the confix *-ποιώ* are used for the formation of denominal or deadjectival verbs in Modern Greek language. Some of them are synchronically productive creating neologisms (e.g. *-ίζω, -(ι)άζω, -εύω, -ώνω, -άρω, -ποιώ: γραμματικοποιώ* 'grammaticalize', *σχετικοποιώ* 'relativize', *ζουμάρω* 'to zoom', *κοπιάρω* 'copy', *γαντζώνω* 'to hook', *στελεχώνω* 'staff', *στοχεύω* 'target', *παπουδιάζω* 'become like an old person', *μαϊμουδίζω* 'imitate monkey's behavior') while others have ceased creating words (e.g. *-αίνω, -ύνω*: *γλυκαίνω* 'sweeten', *δασύνω* 'aspirate').

Nooj's annotations can represent suffixes described by derivational rules or productive morphological grammars. The derived forms inherit inflection from their suffixes.

The purpose of the present study is the automatic recognition of Greek suffixed verbs, in order to ameliorate text annotation of the Greek Nooj module. Meaning assignment to annotated words based on base-suffix semantic combination criteria will also be attempted.

In the first part of the present paper emphasis will be given to the morphological and semantic analysis of the above mentioned suffixes and the derived verbs according to 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 (Classes d'objets Gross 1992) of bases and given suffixes will follow. Finally, a demonstration of derivational rules and productive morphological grammars constructed for the automatic recognition of Greek derived verbs will close this paper.

References

Corbin, D. (1987), *Morphologie dérivationelle et structuration du lexique*, 2 vol., Tubingen.

Corbin, ***D***. (***1991***). « Introduction: la formation des mots-structures et interprétations». *Lexique* 10, 7-. 30.

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

Jackendoff, R. (1990), *Semantic Structures*, Cambridge, MIT Press.

Named entities in Chinese

Ying Yang, Gordana Pavlović-Lažetić, Miloš Utvić, Duško Vitas

HLTG, Faculty of Mathematics, University of Belgrade
violet.yang27 at gmail.com,
(gordana | misko | vitas) at matf.bg.ac.rs

The paper presents resources developed in the Nooj system framework, aimed at recognition of names entities in Chinese texts that use simplified Chinese characters, as well as development of lexically-based text alignment methods. Nooj already possesses resources for Chinese texts that use traditional Chinese characters (produced by Huei-Chi Lin, Université de Franche-Comté). Traditional and simplified Chinese characters are the two standard sets of printed Chinese characters.

The text that we experimented with was a version of Jules Verne's "Around the World in Eighty Days" in Chinese and English. The text is available in electronic form in a number of languages, and within our group it has been aligned in 18 languages, including Chinese. All the texts are in XML format, in accordance with TEI recommendations. Alignment has been produced using the system ACIDE (Aligned Corpora Integrated Development Environment), which integrates XAlign and Concordancier tools from the Loria laboratory, providing for different transformations of parallel texts (including generating TMX and HTML formats of aligned texts). Alignment has been done in such a way that a 1:1 correspondence has been established between English and Chinese segments, providing for Chinese text to be matched with translation in other languages, such as Serbian, Bulgarian, or Hungarian.

On the other side, for Chinese version of the text we developed NooJ-dictionaries that allow recognition of named entities (proper names, date components, measures, etc). By using these dictionaries we constructed graphs recognizing specific classes of named entities (proper names, dates, currencies, distances). Results obtained have been compared with those obtained for other languages (especially for English).

By using BiTMark system, developed in our group, matching of recognized named entities in Chinese and English has been performed and a corresponding bilingual dictionary produced. Main purpose of the dictionary is to provide a resource for improvement performances of the alignment program.

Keywords: Chinese language, NooJ, Named Entity Recognition

Portabilité du corpus arménien sous NooJ

Anaïd Donabédian

Victoria Khurshudyan

Une collaboration étroite a été initiée depuis 2009 entre le corpus d'arménien en ligne http://eanc.net et le projet de corpus arménien sous NooJ conduit à l'Inalco depuis quelques années.

Le corpus en ligne concerne actuellement l'arménien oriental, alors que le corpus NooJ concerne l'arménien occidental. Pour les deux variantes, le lexique diffère fondamentalement peu (environ 5%, le plus souvent la différence est stylistique mais les mots ne sont pas absents du lexique) si on fait abstraction de la question de l'orthographe (l'Arménie, dont la langue standard est l'arménien oriental, a adopté une orthographe simplifiée pendant l'époque soviétique, mais pas la diaspora qui a conservé l'orthographe traditionnelle).

L'objectif de la collaboration est de :

- rendre disponible les textes en ligne à un large public dans l'interface simplifiée élaborée par http://eanc.net
- rendre disponible le corpus d'arménien oriental comme occidental sous NooJ pour les linguistes afin de permettre des requêtes complexes
- utiliser les grammaires syntaxiques de NooJ pour permettre de résoudre les ambiguïtés dans le corpus existant sans traitement manuel
- dans les deux interfaces : permettre des recherches sur les deux variantes

Les défis sont nombreux :

Au plan linguistique :

- rendre compatible les orthographes (plusieurs projets de convertisseurs existent actuellement)
- compiler les dictionnaires, récupérer et filtrer les étiquettes
- modéliser les différences morphologiques entre les deux variantes (environ 20%), vérifier la compatibilité des catégories (les inventaires de formes ne sont pas bilatéraux)
- rendre compatible les étiquettes grammaticales des deux projets (choix terminologiques et théoriques)

Au plan du traitement informatique :

- modifier automatiquement les étiquettes utilisées jusqu'à présent dans NooJ grâce à la fonction recherche-remplace des grammaires graphiques

- mettre à jour les dictionnaires et grammaires NooJ après compilation des deux dictionnaires
- utiliser la portabilité xml pour transférer les textes étiquetés syntaxiquement d'un projet dans l'autre, et pour tester de façon croisée les fichiers

Dans cette communication, nous présenterons l'avancement du projet (une version test du corpus d'arménien occidental est prévue en ligne pour juin 2010).

Cette expérience de portabilité entre l'interface NooJ et une interface grand public peut intéresser les utilisateurs de NooJ sur d'autres langues.

NooJ disambiguation local grammars for Arabic broken plurals

(1)Samira ELLOUZE, (2)Kais HADDAR et (3)Abdelhamid ABDELWAHED
(1) MIRACL, FSEGS, Sfax 3018, Tunisia, ellsamig@yahoo.fr
(2) MIRACL, FSS, Sfax 3000, Tunisia, Kais.haddar@fss.rnu.tn
(3) ELSCA, FLSHS, Sfax 3018, Tunisia, Abdelhamid.abdelwahed@yahoo.fr

One of the problems encountered by researchers in the domain of NLP is without doubt the ambiguities produced by the linguistic applications such as morphological analysis, parsing, etc. In case of ambiguity, we face a choice between two or more alternatives for an analysis phase. The problem resides in choosing the most appropriate alternative.

The detection of broken plural for simple or compound 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 morphosyntactic phenomena related to the Arabic language as the lack of vowels, agglutination, etc. Although the lack of vowels always presents a main source of ambiguity, agglutinative forms may play a dual role: a source of ambiguity or the opposite. In fact, this phenomenon is a source of disambiguation because some enclitic or proclitics can only be attached to noun or adjective.

The disambiguation is a crucial step in the detection of broken plural. We find in literature two main approaches for disambiguation: probabilistic approach and constraints approach. In our work, we have resorted to the constraints approach. This approach allows the development of a list of rules to remove ambiguities. Those rules are classified into three categories: context rules, heuristic rules and non-contextual rules. In each of these three categories of rules, we find the rules for detecting compound nouns in plural broken. We identified several rules to detect this type of plural. These rules are based on different grammatical categories as the compound annex (_____ إ _ آ), the compound replacement (____ _ آ), the compound description (_____ _ آ), etc.

Our work aims reducing the number of interpretations from the detection of Arabic broken plural. Besides, it aims to detect correctly plural broken of compound nouns. And also, the construction of dictionaries and grammars required for detecting the broken plural of simple or compound nouns, with less ambiguity.

To make this work, we first build a corpus for a well specified domain. The choice of the domain is due to the infinite number of compound nouns in plural broken that are found in the Arabic language. Then, we group the enclitic and proclitics: those used only with nouns and adjectives in one group and another in a second group. Then, we establish the list of rules where we encountered a simple noun or a compound noun. Afterwards, we search the additional rules that may apply when there is a broken plural noun only by excluding the other two types of plural (regular masculine plural, regular plural feminine). Finally, we translate the different rules emerged from our study in a set of NooJ transducers, with the addition of needed dictionaries.

Keywords: Local syntactic grammar, compound words, NooJ.

NooJ in Russian Language Content Analysis of Crew Communication of the Mars 105 Space Analog Simulation Experiment

Gushin, Vadim (1), Ehmann, Bea (2), Shved, Dmitri (1), Balázs, László (2)
(1) Institute for Biomedical Problems of the Russian Academy of Sciences,
vgushin@ibmp.ru, sdm84@rambler.ru
(2)Institute for Psychology of the Hungarian Academy of Sciences, ehmannb@mtapi.hu,
balazs@cogpsyphy.hu

The Mars-105 experiment was executed March – July 2009 in Moscow, IBMP with participation of the European Space Agency (ESA). The crew of 4 Russians and 2 Europeans (German and French) spent 105 days in hermetic modules, accomplishing a sophisticated scientific Protocol. During the isolation period, the crew had the opportunity to contact Mission Control (MC) via telephone and computer (written daily reports and E-mail). The influence of isolation and autonomous environment on crew communication content is the subject of our study.

Analysis of crew communication with MC is the standard medical monitoring operation, approved by RSA (Myasnikov et al., 1982). In Mars–105, daily written reports to the MC were content analyzed by both expert assessments and automatically by NooJ software, the latter in cooperation with Hungarian team.

Written crew reports in Russian were processed with NooJ (Silberztein, 2003), which allowed making objective quantitative content analysis. First, expert-defined word categories were made:

- Activity – statements, containing words that express accomplishment of Mission Protocol;
- Needs – statements, containing words that express crew requests and demands;
- Social regulation – statements, containing words that express crew attitudes to the supervisors;
- Positivity/Negation – statements, containing words that express the crew's assessment of their life and work.

Then annotated NooJ dictionaries and searching graphs were made for each category. Using the Corpus Building and Statistics function, the matches were plotted along advancing time in the Mission.

The results showed good concordance with subjective content analysis, made by the group of experts (psychologists) according to the standard procedure (Kanas N., Yusupova A, et al, 2009).

Combining psychological expert knowledge with computerized content analysis in distant monitoring has several advantages. The process is repeatable and controllable; requires no surplus effort or resources; indirect linguistic markers can be used; category making is quick and flexible; the results can be processed by statistical programs, and can be integrated with accumulating psychological expert knowledge on the dynamics of isolated small groups, and the analysis is not bound to English language texts.

Considering the English language dominance of content analytic software available on the scientific market, there is high need of a multilingual linguistic development environment for the psychological content analysis of verbal behavior of multinational crews that is able to produce comparable results. The present Russian language analysis is a step toward the goal of close telemetric monitoring of the psychodynamic events of international crews in several languages involved in space psychology.

Keywords: NooJ, Mars-105, Content Analysis

A Corpus Based Nooj Module for Turkish

Ümit MERSİNLİ(1), Mustafa AKSAN(2)
(1) Mersin University, Turkey, umitmersinli@gmail.com
(2) Mersin University, Turkey, mustaksan@gmail.com

This paper presents the design, implementation and testing processes of a corpus-driven Noojmodule for Part of Speech and morphological tagging of Turkish. The study will illustrate cases that pose specific morphological challenges in analyzing word structure in Turkish. Modeling and tagging processes involve both inflectional and derivational paradigms of present-day Turkish. Multi-word units and syntactic disambiguation of annotated texts are beyond the scope of the study.

Data of the study are derived from the Turkish National Corpus (TNC). The corpus targets a size of 50 million words and follows the construction procedures of BNC. Currently, the corpus contains about 30 million words extending over nine different domains. Data for tokenization are extracted from a sub-corpus including over 100 texts representing different genres taken from TNC. The sub-corpus included over 3,300,000 words forms and over 280,000 tokens when proper nouns, abbreviations and acronyms are excluded.

In the first section of this paper, the principles in modeling and tagging of Turkish as a phonology-driven and rich-morphology agglutinative language will be presented. While order of suffixes in Turkish is relatively fixed, the ordering possibilities and homophoneous nature of roots and suffixes themselves pose serious challenges.

The tokenization and lemmatization processes will be described, including the phonological alternations of Turkish and their applications during the compilation of the dictionaries. Parts of Speech tags used in the dictionaries will also be listed and exemplified.

Followingly, the modeling process, including the suffix tagset used in Nooj morphological grammars designed for Turkish will be presented. Previous work on the listing of Turkish suffixes will be discussed and the problems in the identification of derivational and inflectional suffixes will be presented; homophones, as the main challenge in ambiguity resolution and modeling of Turkish morphology, will also be discussed.

In sum, this paper will discuss the challenges in the design and implementation of Nooj module for Turkish and its testing on various sub-corpora again extracted from TNC. We will argue that among the five types of linguistic resources mentioned in Blanco and Silberztein (2008), Nooj Turkish module will be an example for extended usage of Nooj's morphological parser. The study will conclude with suggestions concerning the future work to improve the accuracy of the module.

Keywords: Turkish, Nooj, corpus linguistics, agglutinative morphology, POS tagging, grammatical tagging, Turkish National Corpus.

References:

Blanco, X. Silberztein, M. ed. (2008). *Proceedings of the 2007 International Nooj Conference*. Cambridge Scholars Press.

Using NooJ in a process of economic intelligence: Producing knowledge for information monitoring and re-indexing content

Philippe LAMBERT, Sahbi SIDHOM
Nancy Université, philippe.lambert@vinalor.fr, Sahbi.Sidhom@loria.fr

In France, more than 15 million people suffer from chronic diseases, their sustainable and scalable, cause disability, personal problems, professional and social consequences. Franchise term care, long-term illnesses.... Given the lack of health insurance that can not be suppressed, governments are trying to find the best solutions to a dilemma: controlling rising health care costs while preserving the sustainability of an effective care system. France founded his health system on universal character. Over time, new issues underlying this problem have been grafted to a purely financial logic: (i) why do health spending increase everywhere in the world? And (ii) why faster in some countries?

It is in this context has been conceived and directed "ChroniSanté », an information system for decision support as part of a Masters thesis on Scientific and Technical Information and Economic Intelligence, combining facets of R & D professional context of INIST in France (ie. Institute of Scientific and Technical Information - Unit of CNRS for the provision of specialized information). ChroniSanté system aimed to help a multidisciplinary working group on chronic diseases of the High Council of Public Health (HCSP) to issue a series of recommendations on reforming the healthcare system in France. This project was developed in partnership between the HCSP and INIST.

Three lines of research with operational purposes motivated the use of NooJ for this work : (*i*) extraction of information for re-indexing of document content (ie bibliographic descriptions enriched content type indexer and author) (*ii*) the visualization of scientific and technical information information to observe hidden knowledge which is important in the context and especially in a process of economic intelligence, (*iii*) the provision of an "cognitive" tool for economic intelligence actors (decision maker or expert in intelligence analyst) for easy reading summary and link between resources (i.e. information), concepts (i.e. knowledge) and semantic relationships (i.e. strong or weak signal between domain knowledge).

The first part of this paper will present the methodology used in our work. Initially, we will propose a definitional framework of our approach to better understand the concepts of intelligence economic and information visualization.

Then, in a second step, we will present the tools we have implemented in NooJ. Starting from a working Sahbi Sidhom to achieve a platform of morpho-syntactic analysis for automatic indexing and information retrieval, we have adapted the language resources of this work for NooJ in order to extract noun phrases (NP) of a textual corpus constituted by several hundreds of bibliographical references. The results of NooJ SN automaton were then subjected to a visualization tool (NodeXL), to identify the major themes of the corpus and the overlap with the keywords equation used in three scientific and technical information databases.

Finally, the last part of our presentation will discuss results and the possibilities and prospects offered by our work.

Mots-clés: Information Extraction, economic intelligence, Natural Language Processing, Data Visualization, Extraction d'informations, intelligence économique, Visualisation de données

Refernces :

[1] Ph. Lambert, S. Sidhom, *Knowledge Extraction and Vizualisation : case study on ChroniSanté project in France*, SIIE, Sousse, Tunisia, 2010.

[2] S. Sidhom, *Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information: de l'écrit vers la gestion des connaissances* , Lyon, 2002.

[3] B. Dousset, T. Dkaki, et J. Mothe, *Veille Scientifique et Technique sur InterNet*, Oct. 2009.

[4] S. Card, J. Mackinlay, et B. Shneiderman, *Readings in Information Visualization: Using Vision to Think. 1999*, Morgan Kaufmann.

NooJ as a Tool for Psychological Content Analysis of Small Group Communication in Isolated, Confined and Extreme (ICE) Environment

Ehmann, Bea(1), Balázs, László(2), Fülöp, Éva(3), Hargitai, Rita(4) and László, János(5)
(1) (2) (3) (5)Institute for Psychology of the Hungarian Academy of Sciences,
ehmannb@mtapi.hu, fulop81@gmail.com, balazs@cogpsyphy.hu
(4) (5) Institute of Psychology, University of Pécs,
hargitairita@freemail.hu, laszlo@mtapi.hu

NooJ (Silberztein, 2003) has been comprehensively used for psychological research in Hungary in cooperation with corpus linguists for several years. Based on the theory of Scientific Narrative Psychology, the method of investigation is termed Narrative Psychological Content Analysis (László, 2008). This new paradigm allows for the description of mental states and representations from human verbal behavior in natural settings. Its results may be predictive to psychological issues and social maladaptation at both individual and group level.

The principle of using NooJ for psychological content analysis is twofold. One method is to use linguistically annotated large Hungarian dictionaries supplied by corpus linguists (Tamás Váradi and Kata Gábor). This forms the basis of creating composite syntactical graphs for searching various psychosemantical matches in texts, such as, among others, Emotion (Éva Fülöp), Self Reference and We Reference (Rita Hargitai). Termed Psychosemantical Modules, these large graphs can be used for content analysis of a great variety of psychological texts, including interviews, self-narratives, etc. The second method is when psychologists themselves create psychosemantically annotated mini-dictionaries and mini-graphs for specific purposes in a particular project.

The present paper intends to demonstrate the use of NooJ as a content analytic tool in the psychological status monitoring of a small group in a space analog simulation environment, the Mars Desert Research Station (MDRS), Utah, USA. Crew 71 was six Hungarian volunteers, who worked there from April 13 to April 26, 2008. The Mission lasted for 13 days; the subjects wrote individual diaries on a daily basis.

The large NooJ Modules were used for the quantitative assessment of Emotional Status (the rate of positive and negative emotions mentioned in the diaries), Group Cohesion (the rate of self and we references mentioned in the diaries), and Subjective Physical Comfort (the rate of physical comfort and discomfort events mentioned in the diaries). Plotted along the advancing time of the Mission, the results showed the day-to-day patterns of the three measures. The data well illustrated the difficult days when the group was in lower mood, was more sensitive to minor physical discomfort, and the members focused on themselves rather than to their friends.

The single purpose NooJ dictionaries and graphs were used for sociometric analysis: the aim was to count how many times each subject was mentioned by others in the diaries. All name variants collected from the token list were compiled to a mini-dictionary, and mini-graphs were made to find the matches. The results showed which crewmember was in the focus of group attention and who was at the periphery on day-to-day basis.

The Corpus Building function and the statistical function of NooJ were indispensable in both analyses. The results were plotted in Excel diagrams.

This pilot study intends to demonstrate that NooJ is an excellent tool for psychological content analysis; in particular, it may be applied in telemetric psychological monitoring of small group communication in ICE environments, such as ships, submarines, military missions and space analog simulations with multinational crew.

Keywords: NooJ, Content Analysis, ICE Environments, Narrative Psychology

Selection criteria for method of translation and some suggestions for the platform NooJ

Hajer SAHNOUN (1), Kais HADDAR (2)
(1) MIRACL, sahnounahajer@yahoo.fr
(2) MIRACL, kais.haddar@fss.rnu.tn

Automatic translation (AT) is a very ambitious domain of research. Researchers are always trying to invent, improve and provide additions to this domain. They fight obstacles in order to resolve problems encountered. These problems are greatly related to the choice of suitable method for a defined translation situation. Besides, they are related to difficulty of treatment of multiple linguistic facts and depth of analysis provided by various translators.

During implementation, designers of translators can make mistakes in choosing the appropriate linguistic method. In fact, this choice should not be arbitrary. Choosing the most suitable linguistic method avoids the providing of an extra effort useless for translation. Also, it avoids superficial effort for requiring applications in terms of depth analysis and abstraction.

Consequently, in this work we propose a number of criteria that influence the choice of the linguistic method of AT. Indeed, the domain which covers the application of translation influence greatly the choice of method. In fact, we judge this criterion of two viewpoints: the generality and the size of lexicon. In addition, the type and the complexity of the structure can lead to the use of a well-defined method. Moreover, among the criteria that must be taken into consideration when designing a translator, the number of languages taken into account by the system. Also, each user has different needs according to his knowledge of the language and the type of activity requiring translation. In fact, there are users who use the automatic translator for identifying relevant documents in the source language, classify documents according to their subjects and understand the overall meaning. Thus, requirements differ greatly depending on the desired use.

Judging the influence of each criterion, independently of others, on the choice of the appropriate method of translation is insufficient and can produce a wrong choice. In fact, it is necessary to take into consideration all the criteria to avoid contradictions. However, taking into account a number of criteria do not solve entirely the problem because some conflicts may appear. Therefore, our proposal is to condition the use of these criteria and to assign them priorities in order to develop an approach facilitating the choice of appropriate linguistic method.

After proposing an approach to choose the proper method of translation, a phase of experimentation and validation is performed with the linguistic platform NooJ. This phase allows us to offer some suggestions for NooJ. Indeed, our observations showed that the majority of works of AT achieved using this platform uses the semi direct method. This does not encourage developers of automatic translators to use it. In this context, we try to deepen the analysis in order to show the possibility of achieving more complex methods (e.g., transfer method). The idea is to exploit analysis works (e.g., parser construction) using NooJ for different languages. Indeed, we are studying techniques used in analysis modules in order to facilitate their reuse in the process of structural and lexical transfer. Moreover, this can help to deepen the analysis during the translation process. We suggest further changes that could improve the current process of translation using NooJ. In fact, these changes aim to increase performance of NooJ in the domain of AT.

Keywords: selection criteria, AT, structural transfer, lexical transfer.

NooJ Module for Sentiment Identification in Croatian Financial Texts

Željko Agić (1), Nikola Ljubešić (1), Marko Tadić (2)
(1) Department of Information Sciences
(2) Department of Linguistics
University of Zagreb
{zagic, nljubesi, mtadic}@ffzg.hr

In this contribution, we present a NooJ module for sentiment identification and analysis in Croatian newspaper texts from the financial or business domain. The module itself is the end result of a previously conducted experiment investigating possible correlations between the Zagreb Stock Exchange index CROBEX movement and the general sentiment presented by the media over certain intervals of time. The experiment has shown that positive correlation does exist between (1) periods of obvious trend (positive or negative) on the stock exchande and general sentiment appearing in newspaper texts and (2) general sentiment encoded in these texts and certain sentiment-denoting phrases found within these texts. It was also shown that expressions of the same sentiment are propagated through neighboring paragraphs of the text. Experiments were made on a Croatian business newspaper business.hr text collection, which was manually annotated for overall sentiment and sentiment phrases. The sentiment phrases – tagged as having either positive or negative weigth, i.e. sentiment – were extracted into frequency lists of positive and negative expressions, which in turn served as a basic input for the construction of the NooJ module. Namely, the frequency lists served to construct cascades of local grammars used to identify business- or finance-related key phrases carrying sentiment information within unseen Croatian texts. These local grammars output two types of metadata regarding the detected phrases: the sign (plus or minus, i.e. positive or negative) of the expression and its empirically assigned sentiment weigth. Basically, we try to address the fact that different sentiment-denoting expressions may vary in the degree or magnitude of the sentiment. For example, *oštar pad dioničkog indeksa* (en. *a steep decrease of the stock market index*) might intuitively be considered to be more negative than *blagi pad* (en. *a slight decrease*). Therefore, in addition to being classified as negative sentiment expressions, these should differ in the amount of negativity they encode and consequently transfer to the end-user, i.e. the news consumer. The module is developed exclusively in NooJ by using local regular grammars, i.e. finite state transducers and Croatian language resources for NooJ. We give insight on the module design, provide its evaluation on unseen Croatian texts and discuss possibilities of integrating the module into larger sentiment analysis systems.

Keywords: sentiment analysis, sentiment identification, financial texts, Croatian language, NooJ

Building a Sanskrit module in NooJ: Basic resources

Vanja Štefanec, University of Zagreb, vstefane@ffzg.hr

Even though the first attempts to use the computer in Sanskrit studies were made already in the 1970s, the interest for computational processing of Sanskrit language increased only ten years ago. By now, a large number of linguistic resources have been developed and made available online, as well as the large corpora of digitalized Sanskrit texts. As to my 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 BC. 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 very important language for Indo-European comparative linguistics, this module should not be interesting only to linguists but to Indian philologists as well. In my paper, I will present some basic resources for Sanskrit module: outline of properties, basic dictionary, few inflectional and derivational grammars and three demo texts with samples of three major styles (epic, literary and scientific).

Keywords: Sanskrit, inflection, derivation, NooJ.

Towards Parsing Croatian Complex Sentences: Dependent Noun Clauses

Vanja Štefanec, Kristina Vučković, Zdravko Dovedan
University of Zagreb,
vstefane@ffzg.hr, kvuckovi@ffzg.hr, zdovedan@ffzg.hr

In this paper, authors will present methods for parsing Croatian complex sentences in which a dependent clause serves as a direct object to the main verb. This research is based on the resources that have already been developed for parsing simple Croatian sentences.

So far, sentences that we were able to parse using these resources are of the basic structure consisting of a subject, verb, direct and indirect object, adverbial of time and place. Methods we shall present in this paper will extend this structure to the following sentence structure **<main clause** <*dependent clause*>> and, although quite rare and stylistically marked, to the structure <<*dependent clause*> **main 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).

Since this type of complex sentences is very common in Croatian language, we believe that this research will be a valuable contribution to Croatian module for NooJ. At the end of the paper, we will evaluate the adequacy of the model through precision, recall and fmeasure.

Keywords: parse trees, partial parser, main clause, dependent clause, Croatian language, NooJ.

Arabic Compound Nouns Processing: Inflexion and tokenization

Ines Boujelbene (1), Slim Mesfar (2), Abdelmajid Ben Hamadou (1)
(1) MIRACL, ISIMS, Univesity of Sfax, Tunisia
boujelben_ines@yahoo.fr, abdelmajid.benhamadou@isimsf.rnu.tn
(2) RIADI, University of Manouba, Tunisia mesfarslim@yahoo.fr

In this work, we deal with the construction of compound nouns[1] linguistic resources in order to improve the lexical coverage of our biomedical terminology extractor. We start with the extraction of the compound nouns by means of some NooJ’s regular expressions. These regular expressions are applied to our specialized corpus containing about 1400 medical texts and including over one million word forms. This first extraction step gives more than 2000 compound nouns related to the biomedical domain. Then, we classified the list of identified expressions into 25 categories divided among:

- 9 categories for compound words formed by the succession of two word forms
- 13 categories for compound words formed by the succession of three word forms
- 3 categories for compound words formed by the succession of four word forms

After that, since the Arabic is a highly inflectional and an agglutinative language, we were faced to the problem of generating all the potential inflected forms as well as the recognition of agglutinated forms. In fact, commonly, we have either inflected forms or agglutinated ones rather that the bare form of lexical entries. Thus, in the first hand, we tried to deal with the compound noun inflection using the traditional methods based on the writing of inflectional and derivational paradigms but we noticed that is would require more than 500 paradigms. In the other hand, we observed 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. Thus, we suggest solving these problems by means of a set of syntactic grammars. The proposed grammars would recognize all utterances of inflected compound nouns without saving them into an already compiled dictionary. They, also, recognize all the agglutinated forms where conjunctions, prepositions and personal pronouns could be added to the initial lexical entry. In addition, the built local grammars are able to inherit the whole available information (POS, inflectional, distributional and semantic information) to annotate the recognized compound noun.

For instance, these grammars are able to recognize, lemmatize and annotate the two following expressions:

نَوْ ا بَ ت قَلْبِيَّة (nawbaat qalbiyyah – heart attack**s**, in the plural) : an inflected form

وَ ن بِ وْ ت بَ ه القَلْبِيَّة (wabinawbatihi elqalbiyyah – **and with his** heart attack) : an agglutinated form using the lexical entry:

نَوْبَة قَلْبِيَّة ,N+N_ADJ[2] +Maladie+Cardio[3] (nawbah balbiyyah – a heart attack).

Automatic acquisition of Bulgarian FrameNet candidates with NooJ

Svetla Koeva, Department of Computational Linguistics, Institute for Bulgarian,
svetla@dcl.bas.bg

In this paper we describe series of NooJ local grammars used for an automatic enlargement of the Bulgarian FrameNet with new verb senses. The first Section shortly presents the structure of the Bulgarian FrameNet. The frame lexicon entry at the Bulgarian FrameNet consists of a target verb, its unique definition, annotated examples, corresponding semantic frame from the English Framenet, adjacent grammatical class (values of the attributes person, transitivity and aspect) and syntactic frame (a set of feasible syntactic structures associated with the target verb). The syntactic structure defines the number of arguments uniquely specified for a syntactic category (with particular sets of prepositions or complementizers, if any), lexical explicitness, grammatical function, and four sets of semantic restrictions (corresponding to the respective top most senses from the Bulgarian WordNet).

In Section two, we discuss the Bulgarian language specific features that cause difficulties in formulating local grammars for an automatic acquisition of verb senses and their classification to the predefined verb frames. The most prominent among them are the relatively free word order, the null subject and possible implicit

[1] A compound noun is a composition of a new word by the addition of two or more simple forms and having a unique meaning.

[2] N_ADJ : a compound noun composed by a noun followed by an adjective

[3] +Maladie+Cardio : is a distributional information

realization of the rest of the arguments, the high morphological and syntactical ambiguity, the recursion, the ambiguity of the constituent's scope and so on. At the same time we explain why the task to construct local grammars resolving all problems mentioned is too complicated.

In Section three, we describe the argument structure distinctions of the Bulgarian verb classes in more details. Verb frames can be classified according to the particular sets of features chosen from the information encoded so far. For example if we consider only the number of arguments, their categories, their explicitness, and the sets of possible prepositions the quantity of verb frames will be significantly reduced. Thus the task of automatic acquisition of verb senses belonging to different verb frames relates to the question what information about verbs and their arguments is specific enough to allow their automatic classification. In other words, we limited our work to the construction of local grammars for those verb frames that show unique unambiguous features, i.e. obligatory explicit realization of a specific type of arguments (i.e. verb clitics); inanimateness of the subject combined with the limited verb paradigm and so on.

In Section four, series of NooJ local grammars automatically classifying verbs based on the specific argument structure properties, learning experiments with them over a large corpora and a detailed analysis of detected errors, are presented. As a result we show that local grammars formulating the distribution of the unique features are successfully used for the acquisition of the verb sense candidates for respective verb frames. Section five evaluates the significance of these results by comparing the local grammar's accuracy to an expert based classification.

We conclude the paper with a discussion of its contributions, comparison to related work, and suggestions for future extensions. The idea to search for new verb senses belonging to one and the same verb frame has been exploited for some languages but to the best of our knowledge Bulgarian is not among them.

Bulgarian National Corpus Project

Svetla Koeva, Diana Blagoeva, Siya Kolkovska
Institute for Bulgarian, svetla@dcl.bas.bg, diablag@mail.bg, sia_btb@yahoo.com

National electronic corpora are representative of the state of a particular language at a certain period of its development. During the last decades, national electronic corpora were developed for most of the Slavic languages - Czech, Slovak, Polish, Croatian, and Russian. The National Corpus of Bulgarian - BulNC, (Български национален корпус) is another large Slavic corpus project.

The Bulgarian national corpus project is building a large-scale, representative, publicly available corpus of Bulgarian. The BulNC can also be defined as a monolingual general corpus, fully annotated morphosyntactically (and partially semantically). Presently the Bulgarian National corpus consists of about 320 000 000 words and includes more than 10 000 texts. All materials in the corpus are drawn from written materials (and a few transcripts of spoken data) produced after 1945. The BulNC incorporates four general sub-corpora provided with a uniform metadata description and morphosyntactic annotation, which facilitates their processing and grouping according to different criteria. The individual parts of the BulNC are provided with detailed metadata description in a unified XML format.

The whole corpus is annotated (for parts of speech, detailed grammatical information and Bulgarian WordNet word senses). Because of the size of the corpus, only parts of it are manually annotated: + 300 000 words for parts of speech and + 150 000 words for word senses. In various kinds of annotation we follow established standards to the extent that they are compatible with the language specific features of Bulgarian - we share the understanding that the standardization in morphosyntactic annotation has to cover both correspondences between different languages as well as language specific features. For this purpose the NooJ Annotation System for corpus processing is very appropriate, because it provides a number of parsers that operate in a cascade way. In particular we implement the Bulgarian Nooj dictionary of simple and compound words (+85 000 entries) as an inflectional analyzer, NooJ's local grammars for the annotation of certain word sequences: multi-word units and semi-frozen expressions. Moreover, NooJ allows multiple annotations resulting in complex embedded annotations as well as adding and removing annotations at the same time.

The Bulgarian National Corpus is an extremely valuable electronic resource not only for Bulgarian and foreign linguists, but also for a wider range of users: humanitarians, teachers, students, translators and anyone who is interested in the state and the development of Bulgarian during the last half century. The work on the enlargement of the BulNC with respect to its better representativeness and well balancing is continuing as well as towards the sophisticated complex annotation using NooJ Annotation System.

Verbal Prefix in Polish: derivation, aspect and semantics

Ewa Gwiazdecka
ASPEKTY Foundation for development of computational linguistics, logic and researches on language,
ewa.gwiazdecka@gmail.com

Verbal prefix in Polish language encodes perfective aspect, but being historically related to spatial preposition it also introduces semantic changes. In Polish, 17 prefixes of Slavic origins can be used to build a new verb in the following way: *pisać*$^{\text{IMPF}}$'to write' – ***na****pisać*$^{\text{PERF}}$ 'to achieve writing', ***do****pisać*$^{\text{PERF}}$ 'to finish writing', ***od****pisać*$^{\text{PERF}}$ 'to write back, to copy', ***po****pisać*$^{\text{PERF}}$ 'to write for a while', ***pod****pisać*$^{\text{PERF}}$ 'to sign', etc. The description of the compositionality between the prefix and the verb is problematic as it implies not only the signification of the two units, but also the aspectuality. Thus, some linguists explain this process by providing the set of inherent properties of the predicate and by building the aspectual verb classes (see, for example, Antinucci and Gebert 1967, Cockiewicz 1995, or more recently, the work done by Młynarczyk 2004).

In our contribution, we will present a series of derivational grammars for prefixation based on the verb classification. This solution will restrict the productivity avoiding the recognition of the incorrect forms, like ****do**kochać* or ****prze**rozumieć*. The aspectual verb properties will be added to the dictionary and will be explained by the complementary work on this subject.

The second part of our presentation concerns the prefix interchangeable derivation. It is commonly admitted to provide the description of the prefixed verb in relation to its base. In this respect, the verb *łączyć* ‘to connect’ can build the following prefixed forms: *dołączyć*, *odłączyć, połączyć*, *podłączyć*, *przełączyć*, *rozłączyć*, *włączyć*, *wyłączyć*, *załączyć*, *złączyć*... However, it has been observed (Jadacka 2007, Skarżyński 2003) that some prefixes of these derived words can enter into a specific semantic relations (Stankiewicz 2005): ***w**łączyć* ‘to switch on’ -> ***wy**łączyć* ‘to switch off’; ***do**łączyć* ‘to connect; to join (a group) -> ***od**łączyć* ‘to disconnect; to leave (a group)’; *złączyć* ‘to join, to connect, to link’ -> ***roz**łączyć* ‘to separate, to divide, to split’. Moreover, we note the "net" of significations relative to the meaning of several prefixes and to the spatial or temporal phases of the process. For example, for the base gnić ‘to rot’, we can produce: ***pod**gnić* 'to begin rotting' -> ***nad**gnić* 'to start rotting' -> **z**gnić 'to rot completely' -> ***prze**gnić* 'to rot across'.

Based on the semantic classification of the relations expressed by the verbal prefix, we propose the series of derivational grammars to give account of this problem.

Keywords: verbal prefix; derivation; prefix interchangeable derivation; aspect; semantics; verb classes

VIET4NooJ: A Vietnamese module for NooJ

(1)Philippe Lambert, (2) Michel Fournié, (3) Océane Ho-Dinh,
(1)VinaLor, philippe.lambert@vinalor.fr , (2) INALCO, mfourn@inalco.fr, (3) CRIM-INALCO, ho.dinh.oceane@no-log.org

The Vietnamese as we know it today, that is to say with a romanized alphabet dates from the seventeenth century when Alexandre de Rhodes, a Jesuit Avignon published the first books on language of Annam. The Vietnamese is a monosyllabic and polytonal language, with relatively homogeneous regional differences concerning only a few phonemes and lexemes. The history of Vietnam is rich in dynamic contacts with many other languages from different asian civilizational spheres (Chinese, Khmer, Malay, etc..) or more distant ones (French, American, etc..).

Only recently have the Vietnamese been given special attention in the field of natural language processing. Several research projects involving French Vietnamese teams in the programs of scientific cooperation programs have worked on the formalization of the Vietnamese language through TAG formalism (Tree Adjoining Grammar). Starting from a corpus consisting of several thousand Vietnamese news release, our study has two main goals : the first one is to identify and extract Named Entities for the Vietnamese and the second one, to create ressources to automatic extraction of Noun Phrases.

The first part of the presentation will concern the history of the draft constitution of a vietnamese module for NooJ. We will espacially outline technical difficulties encountered in coding Vietnam due to a variety of encoding fonts (Unicode, ASCII, etc..) and the solutions which have been chosen. This point remains problematic in the field of natural processing language for Asian languages. The second part of the presentation will explain in a first time our tagging logic (i.e. the tag we chose) and our positioning compared to the different vietnamese grammar schools. A second time will be devoted to a brief state of the art concept of named entities for the Vietnamese. Then, we present the sample text used to develop our work and the language resources produced in order to make the module available to the Vietnamese community NooJ. The third part will present the results obtained through the use of syntactic graphs to extract named entities and noun phrases from our corpus.

Keywords: vietnamese, asian language, Natural Processing Tool

References :

[1] C.T. Nguyen, T.K. Nguyen, X.H. Phan, L.M. Nguyen, et Q.T. Ha, “Vietnamese word segmentation with CRFs and SVMs: An investigation,” Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 2006), 2006.

[2] T.K. Nguyen, “EXTRACTION DES ENTITÉS NOMMÉES VIETNAMIENNES,” 2007.

[3] T.P. TX, T.Q. Tri, A. Kawazoe, D. Dinh, et N. Collier, “CONSTRUCTION OF VIETNAMESE CORPORA FOR NAMED ENTITY RECOGNITION,” Health, vol. 43, pp. 42–00.

[4] D. Dien et D. Dien, “Cognitive linguistics approach to Vietnamese noun compounds,” Mon-Khmer Studies: A Journal of Southeast Asian Linguistics and Languages, vol. 32, 2002, pp. 145–161.

[5] T. Pham, T. Le, et Q. Do, “Bref aperçu sut l'histoire de l'étude des parties du discours en Vietnamien,” Histoire, epistemologie, langage, vol. 26, 2004, p. 145.

Automatic processing of temporality for VIET4NooJ

Nicolas Boffo, Océane Ho Dinh,
University Montpellier 3, nicolas.boffo@univ-montp3.fr
INALCO, ho.dinh.oceane@no-log.org

As part of the development of VIET4NooJ, a system that will identify events in texts and calculate the temporal relations between them, we have formalized temporality in Vietnamese in terms of lexico-syntactic, semantic, and pragmatic/praxematic aspects.

The automatic processing of temporality in Vietnamese just begun. Therefore, our task is complex due to a lack of NLP, lexical, and semantic resources for Vietnamese. Nevertheless, we have assembled a fairly complete bibliography dealing with temporality in Vietnam (Cao Xuân Hạo, Danh Thành Do-Hurinville, Trần Kim Phượng, Lê Khả Khế and Nguyễn Lân).

Moreover, on the French side the processing of temporality has yielded good results especially with the model of the semantics of the temporality of Laurent Gosselin and the S-language of Sylviane R. Schwer.

With these resources and linguistic research we were able to implement the various steps of processing of temporality.

Therefore, we will present to you the dictionary Vietnamese/French of temporal markers and what we have achevied.

Using this dictionary we were able to construct lists of temporal markers classified by semantic aspectuo-temporal values. These lists are directly involved in the processing of temporality within the rules and allow time to identify events in a text corpus in Vietnamese. We will provide demonstrations of tracking temporal entities. Finally, to formalize the time and aspect in Vietnamese we have created a set of rules (algorithms) implementable in NooJ, that calculate the temporal ordering of events in a Vietnamese text. We will demonstrate the operating time of a few rules.

Eventually we plan to computational implement an automatic translation of temporal relations tool that will have decisive importance for the automatic processing of Vietnamese language (Automatic Translation, Computer Aided Translation, automatic summarization, automatic understanding, finding named entities,…)

Key words: Formalization, temporality, time and aspect, temporal annotation, Natural Language Processing (NLP), automatic translation, Vietnamese language.

References :

Cao Xuân Hạo, 1998, « Về ý nghĩa Thì va Thế trong tiếng Việt » (Time et aspect in Vietnamese), Ngôn ngữ 5, 1-32.

Danh Thành Do-Hurinville, 2009, *Temps, Aspect et Modalité en vietnamien ; Etude contrastive avec le français (Tense, Time, Aspect, and Mood in Vietnamese ; Comparative Studie with French)*, L'harmattan.

Durand A. Irène et Schwer R. Sylviane, "A Tool for Reasoning about Qualitative Temporal Information: the Theory of S-Languages with a Lisp Implementation".

Gosselin Laurent, 1996, *Sémantique de la temporalité en français ; Un modèle calculatoire et cognitif du temps et de l'aspect (Semantics of temporality in French ; A cognitive and computational model of tense, time and aspect)*, Duculot.

Lê Khả Khế et Nguyễn Lân, 2001, *Từ Điển Việt Phấp (Dictionary Vietnamese-French)*, nhà xuất bản vãn hóa sài gòn.

Schwer R. Sylviane, 2009, « Représentation du Temps, relations temporelles et théories des temps verbaux » (Representation of Tense and Time, Temporal relations and theories of verb tenses).
Trần Kim Phượng, 2008, *Ngữ Pháp Tiếng Việt ; Những vấn đè về thời, thể (Vietnamese Grammar/ Problem of Time, Tense and ascpect),* Nhà Xuất Bản Giáo Dục.

Vietnamese classifiers processing for nominal syntagms extraction

Hô Dinh Océane, ho.dinh.oceane@no-log.org

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 for 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, with a state of art of the works made on these words, 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. After a description of the corpus used for the tests, 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 raised 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 justify these choices presenting the results found with the part of our corpus used as reference corpus and the tests made on the rest of the corpus.

We will finish with a discussion on what perspectives remain to be explored to improve the system.

Keywords: Classifiers*,* Nominal Syntagm, Vietnamese, NLP, NooJ, Viet4NooJ

References:

[1] Nguyễn Phú Phong, « Questions de linguistique vietnamienne. Les classificateurs et les déictiques », 1995, Presses de l'Ecole française d'extrême-orient, Paris

[2] Nguyễn Tương Hùng, « The structure of the Vietnamese noun phrase » *(thèse de doctorat),* 2004, Boston University, Boston, USA

[3] Do-Hurinville D. T., « Nominalisation et construction du thème en vietnamien », *Faits de Langues 30, 209-216*, 2008.

[4] Cao Xuân Hao, « Hai loại danh từ của tiếng Việt », *in Tiếng Việt, mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa, 265-328*, 1998, NXB Giáo Dục, Vietnam. 2007, NXB Khoa hoc Xã hoi, Hà Noi

[5] Truong, Văn Chình, « Structure de la langue vietnamienne », 1970, Impr. nationale, P. Geuthner, Paris

[6] Lê Văn Lý , « Sơ thảo ngữ pháp Việt Nam » 1968, Sài Gòn, Vietnam.

[7] Grinevald C., « Typologie des systèmes de classification nominale », *Faits de Langue 14, La catégorisation dans les langues, 101-122*, 1999.

[8] Bisang W., « Classifiers in East and Southeast Asian Languages : counting and beyond », *Changes in Numeral Systems, 113-185*, 1999, Mouton de Gruyter, Berlin.

[9] Allan Keith, « Noun and Countability », *Language, 56, 3, 541-557*, 1980.

[10] Allan Keith, « Classifiers ». *Language, 53, 2, 285-311*, 1977.

[11] Senft, Gunther., « Systems of nominal classification », 2008, Cambridge University Press, Cambridge.

[12] Aikhenvald, Alexandra Y., « Classifiers: A typology of noun categorization devices » *Oxford studies in typology and linguistic theory,* 2000, Oxford University Press, Oxford.

The quiver of the algebraic mathematical models

Kambakis-Vougiouklis, Pinelopi, Vougiouklis Thomas,
Democritus University of Thrace
pekavou@helit.duth.gr

Mathematics has always supplied models of analysis to almost every applied science, including linguistics. These models are mainly borrowed 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. Interestingly those sciences do not appear to have anything to do with mathematics at first sight. Thus, we propose two general models of a natural language construction which come from Algebra but they equally apply to any branch of Mathematics. The First General Model consists of five steps or stages while the Second General Model focuses on the perpetual contrast between quotient and product. The First Model could contribute to reveal parameters which may ‘hide’ in the process and, after been revealed, may come up with a better interpretation of the results and, consequently, to safe conclusions. The Second

Model, on the other hand, could successfully lead to the development and use of the so-called ‘small’ mathematical models. Consequently, we suggest the employment of such models in vocabulary use, independently of process or method, in order that certain hidden parameters might be revealed and result in any improvement of the recording and/or reading. Moreover, the bipole quotient-product is expected to contribute to the continuous organization of knowledge and the vocabulary assimilation by the learners.

Such an application presupposes that all languages of the world belong to one ‘universal’ language and any transformation is expected to reveal interesting invariant elements. These invariant element consist the basis of the structure of the model that the vocabulary user is invited to use.

The In present paper we focus on one of the suggested models, namely the Cartesian product and quotient procedure. The model is analyzed and a number of applications in language teaching and learning with specific examples. The proposed model is within the scope of globalization of sciences; yet, our firm belief is that special characteristics should be preserved and the invariant elements should be consolidated.

Therefore, it is very important that, when mathematical models are used in LT and LL research, extra attention to be paid so that every step should be investigated for a complete development of the model.

In present paper we propose, at least at research level, the application of mathematical models, arising from the general way of development of any branch of Mathematics, in language teaching. Overall, two ways of complete development of prototypes are suggested; however, for the needs of present investigation, we will focus only on the prototype which consists of two “steps”, or “stages”, namely the Cartesian product and the quotient. From these two stages, the Cartesian product seems to be quite simple and what is crucial about it is to be recognized. On the other hand, the quotient seems to be rather complex, it is not unique and, moreover, requires special knowledge, practice and experience so that its potential and utility would be revealed.

In present paper we propose a general model of a natural language construction and we hold that such a model may result from the general way of development of any branch of Mathematics. The different steps of development of any language are still to be defined and 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 with specific examples. The proposed model is within the scope of globalization of science; yet, our firm belief is that special characteristics should be preserved and the invariant should be consolidated.

The different steps of development of any language have been widely investigated within the frame of generative grammar but they are still to be defined and further refined and specified, as language is a natural phenomenon in perpetual evolution. Consequently, the mastery of different steps of this non-stopping process may primarily lead to a better comprehension of the parameters and the potential of the structures and finally allow us to reach reliable conclusions.

Lexical recognition in modern Chinese

Huei-Chi LIN lin_huei_chi@yahoo.fr
Université de Franche-Comté, LASELDI

A major issue for Chinese language research is how to treat lexical units: single units, compound words and multiword lexical units. The linguists have different theories to treat the last two types of lexical units, which often provide the lexical ambiguities. We mention some examples:

理髮 (*lǐfă*) <operate-hair> 'have a haircut'

放假 (*fàngjià*) <put-vacation> 'have vacation'

刀子口 (*dāozi kǒu*) <knife-mouth> 'straight talk'

研究員 (*yánjiùyuán*) <research-K> 'researcher'[1]

翻譯系統 (*fānyì xìtǒng*) <translate-system> 'translation system'

浪漫主義 (*làngmànzhŭyì*) <romantic-SK> 'romanticism'[2]

According to linguistic theory, these examples are not single units. The linguists treat them by discomposing into their components.

However, it is more interesting to treat them as single units in NLP, because an engine user search often (*yánjiùyuán*) 研究員 rather then its components (*yánjiù*) 研究 and (*yuán*)員. This user searching also for (*làngmànzhŭyì*) 浪漫主義, but not for (*làngmàn*) 浪漫 and (*zhŭyì*) 主義. The same logic applies to other examples. In this point of view, it is better to treat these examples as one lexeme (token).

In this paper, we propos some criteria that allow distinguishing lexical units (tokens) from free compound words. According to these criteria, the lexical units like (*fàngjià*) 放假, (*dāozi kǒu*) 刀子口 or (*fānyì xìtǒng*) 翻譯系統 will be treated as one indivisible unit.

Then, we apply these criteria while building Chinese lexical resources. After that, based on these Chinese lexical resources, a lexical segmentor can identify such lexical units in Chinese corpora and do not segment them into their morphological components.

References:

[1] K is the code used in NooJ for Chinese suffixes.

[2] SK is the code used in NooJ for Chinese semi-suffixes.

Guo Rui 郭锐. (1999). Yǔwén cídiǎn de cí xìng biāo zhù wèntí 语文词典的词性标注问题 ‘Problèmes de l’annotation des catégories de mots dans les dictionnaires’. In *Zhōngguó yǔwén 中国语文 ‘Studies of the Chinese Language’*. No. 269. Beijing 北京. pp. 8-25.

Huang Changning et Zhou Qiang 黄昌宁和周强. (1994). Miànxiàng yǔ liào kù biāo zhù de hànyǔ yīcún tǐxì de tàntǎo 面向语料库标注的汉语依存体系的探讨 ‘Approach to the Chinese Dependency Formalism for the Tagging of Corpus’. In *Zhōngwén xìnxí xué bào 中文信息学报 ‘Journal of Chinese Information Processing’*. Vol. 8. No. 3. Beijing 北京. pp. 35-52.

Silberztein Max (a). (2004). NooJ : an oriented object approach. In *INTEX pour la linguistique et le traitement automatique des langues*. Les Cahiers de la Maison des Sciences de l’Homme Ledoux. Presses Universitaires de Franche-Comté, Besançon.

Silberztein Max (b). (2007). An Alternative Approach to Tagging. Invited Paper In Proceedings of 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007: Natrual Language Processing and Information System. Coll. «LNCS series (4592)». Berlin / Heidelberg : Springer-Verlag. pp. 1-11.

Recognition of Names of Libyan Persons Using Nooj Plateform

Abdelsalam Almarimi(1), Hela Fehri(2), Khalid Hussain(1), Abdelmajid BEN HAMADOU (2)
(1), Higher Institute of Electronics Baniwalid, Libya
belgasem_2000@yahoo.com, K1982_2007@yahoo.com
(2), Research Laboratory Miracl, IMS-Sfax, Tunisia
hela.fehri@fss.rnu.tn, Abdelmajid.benhamadou@isimsf.rnu.tn

The objective of this project is to design a system for the recognition of Named Entities on Libyan proper Names in order to translate them into English using NooJ platform. In fact, the Libyan Proper Names have a syntax quite specific and are based on a particular vocabulary (repetition of first names, adding nicknames like (Al-Ostath, Al-Akh, Al-Mohandez, Al-Doctor, Al-Chaikh,..). This article begins by presenting the specificity of the Libyan Proper Names on the level of the vocabulary and the local syntax. Then an implementation of the recognition process using the platform NooJ is given. The developed system will be made available to academic and government institutions to help them for the proper translation of this type of Named Entities.

This system is realized through a cooperation project between the Higher Institute of Electronics BaniWalid, Libya and the research laboratory Miracl, University of Sfax. At present, we have come to constitute a corpus of texts containing this kind of

Named Entities and creating a dictionary of the specific vocabulary. Also, the implementation of local grammars using the Nooj platform is underway. We also plan to define a generic form of internal representation of the recognized Named entities in order to facilitate the translation into English and reuse the recognition process.

Keywords: Named entity, Libyan Proper Names, translation, generic internal representation.

La traduction dans une communauté multilingue - - les outils, les besoins, l'acquis et les problèmes

Mário Vilar.

Avec cette intervention nous voudrions partager avec les autres participants à ce congrès l'expérience concernant l'univers commun aux dictionnaires et grammaires électroniques, aux programmes de traitement de texte, à la terminologie et à la traduction automatique qu'est celle du service de traduction de la Commission européenne (DGT).

Soit dans le souci de la terminologie appropriée, soit dans un souci d'économie de temps, tout en respectant les règles grammaticales et le style propre à chaque langue, la DGT s'est tournée depuis des décennies déjà vers des outils électroniques d'aide à la traduction, qui se sont avérés de plus en plus indispensables, même si pas encore parfaits.

Nous ferons, donc, référence aux dictionnaires, glossaires et bases terminologiques que nous utilisons, ainsi que au système de traduction automatique, au système de traduction liée à des mémoires et au système de reconnaissance vocale, développés spécialement pour nous.

Surtout, nous parlerons des difficultés que avons rencontrées, celles que nous avons surmontées et comment, et les problèmes qui subsistent.

Local grammars for the recognition of negative paraphrases

Angels Catena
Universitat Autònoma de Barcelona
angels.catena@uab.cat

The aim of my communication is to propose a rule-based approach using Nooj in order to identify negative paraphrases in a question answering system.

This research is the result of the collaboration between the researchers at *f*LexSem laboratory (Autonomous University of Barcelona) and INBENTA (http://www.inbenta.com/) that focuses on developing technology for corporate

semantic search. This technology is based on a Semantic Search Engine that provides precise and satisfactory answers to user's queries made in natural language.

My research work aims to achieve the automatic recognition of queries that have the same meaning and are related by negation such as: *préstamo impagado / préstamo que no se ha pagado; carecer de antecedentes penales / no tener antecedentes penales; tarjeta sin registrar / tarjeta que no está registrada; mi sueldo es insuficiente / mi sueldo no es suficiente/ Parece que no hay saldo / No parece que haya saldo…*

We will describe different lexicographical information stored in the database in order to model these lexical relations (using a subset of lexical functions) and propose a set of rules taken from local grammar.