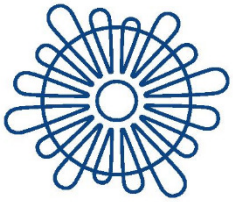

Sveučilište u Zadru
Universitas Studiorum
Jadertina | 1396 | 2002 |

THE 17th NOOJ INTERNATIONAL CONFERENCE 2023

Book of Abstracts

Zadar, May 31 – June 2, 2023

Editors

Linda Mijić
Department of Classical Philology
University of Zadar
Zadar, Croatia

Anita Bartulović
Department of Classical Philology
University of Zadar
Zadar, Croatia

Marijana Tomić
Department of Information Sciences
University of Zadar
Zadar, Croatia

Laura Grzunov
Department of Information Sciences
University of Zadar
Zadar, Croatia

Kristina Kocijan
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia

Max Silberztein
Université de Franche-Comté
Besançon, France

Publisher

University of Zadar
For the Publisher: Dijana Vican, rector

ISBN 978-953-331-421-1

Organization

Organizing Institutions

Department of Classical Philology
Department of Information Sciences
University of Zadar, Croatia

NooJ Association

Centre de Recherches Interdisciplinaires et Transculturelles (Université de Franche-Comté, Besançon)

Organizing Committee

Linda Mijić (University of Zadar, Croatia)
Anita Bartulović (University of Zadar, Croatia)
Marijana Tomić (University of Zadar, Croatia)
Laura Grzunov (University of Zadar, Croatia)
Kristina Kocijan (University of Zagreb, Croatia)
Max Silberztein (Université de Bourgogne Franche-Comté, France)

Scientific Committee

Max Silberztein, *Program Committee Chair* (Université de Bourgogne Franche-Comté, France)
Marco Angster (University of Zadar, Croatia)
Farida Aoughlis (Mouloud Mammeri University, Algeria)
Anabela Barreiro (INESC-ID, Portugal)
Anita Bartulović (University of Zadar, Croatia)
Magali Bigey (Université de Franche-Comté, France)
Xavier Blanco (Autonomous University of Barcelona, Spain)
Krzysztof Bogacki (University of Warsaw, Poland)
Christian Boitet (Université Joseph Fourier, Grenoble, France)

Contents

INVITED TALKS **1**

From Corpora and Corpus Tools to Dictionaries and Beyond: Infrastructure for Slovene

Iztok Kosem 2

Using Semantic Hypergraphs to Represent Knowledge in Natural Language Text

Branko Žitko 4

MORPHOLOGICAL & LEXICAL RESOURCES **6**

Quechua-Spanish and Spanish-Quechua Electronic Dictionaries of Verbs for NLP

Duran Maximiliano 7

A NooJ Dictionary for Italian Light Verb Constructions

Alessia Nicola, Giocondo Cirillo 9

Attenuative Collocations and Parametric Verbs in Old French and Old Spanish: A Contrastive Study

Xavier Blanco, Rafael García Pérez 11

German *Selbst*-Compounds: A NooJ Grammar

Marco Angster 13

Spelling Error Detection and Correction for Arabic Using NooJ

Rafik Kassmi, Samir Mbarki, Abdelaziz Mouloudi 15

Towards a Linguistic Annotation of Arabic Legal Texts: A Multilingual Electronic Dictionary for Arabic

Khadija Ait ElFqih, Maria Pia di Buono, Johanna Monti 17

The Construction of a Multilingual Legal Ontology

Ismahane Kourtin, Samir Mbarki 19

Automatic Translation of Continuous and Fixed Arabic Frozen Expressions Using NooJ Platform

Asmaa Kourtin, Samir Mbarki 21

Recognition of Frozen Expressions in Belarusian NooJ Module

Yauheniya Zianouka, Mikita Suprushuk, David Latyshevich, Yuras Hetsevich 23

A Prototype of Indonesian Multi-Level Tagger: SANTI-Network

Prihantoro 25

NooJ Dictionary for Rromani Language: Importing of a Published Dictionary to the NooJ System

Masako Watabe 27

Deciphering the Nomenclature of Chemical Compounds in NooJ

Kristina Kocijan, Krešimir Šojat, Tomislav Portada 29

Croatian Cognition Verbs in Machine Sentence Processing

Marta Petrak, Bojana Mikelenić, Marko Orešković 31

NooJ Dictionary of Croatian and English Internet Slang

Ivan Cota 33

Latin Pronouns, Numbers and Prepositions in the NooJ Tool

Anita Bartulović, Linda Mijić 35

SYNTACTIC & SEMANTIC RESOURCES 37

Disambiguation Grammars for the Ukrainian Module

Olena Saint-Joanis 38

A Proposal for the Processing of the Nucleus Verb Phrase of Pronominal (SVNPr) Verbs in Spanish

Andrea Rodrigo, Rodolfo Bonino, Silvia Reyes 40

A Rioplatense Spanish Date Grammar Using the NooJ Platform

Mariana González 42

Parafrasário: A Variety-Based Paraphrasary for Portuguese

Anabela Barreiro, Ida Rebelo, Cristina Mota 44

CORPUS LINGUISTICS & DISCOURSE ANALYSIS 46

The Limitations of Training Corpus-Based Methods in NLP

Max Silberztein 47

Syntactic-Semantic Analysis of Perception Verbs in the Croatian Language

Daša Farkaš, Kristina Kocijan 49

The Automatic Translation of Arabic Psychological Verbs Using NooJ Platform

Asmaa Amzali, Mohammed Mourchid 51

Automatic Disambiguation of the Belarusian-Russian Legal Parallel Corpus in NooJ

Valery Varanovich, Mikita Suprunchuk, Yauheniya Zianouka, Yuras Hetsevich, Nastassia Yarash 53

Advances in the Automatic Treatment of Newspaper Articles on Economics Journalism Using NooJ

Carmen González 55

Sentiment Analysis of Texts Written in Arabic: Addressing the Issue of Negation

Mohamed El Ammari, Azeddine Rhazi, Salim Rami 57

A Linguistic Approach for MDU-Based Segmentation

Chahira Lhioui, Malek Lhioui, Mounir Zrigui 59

Embracing a Plant-Based Diet: A NooJ Analysis

Isabella Cossidente, Alessandra D'Agostino 61

Differences Between Hate Comments and Insult Comments Directed to Men and Those Directed Towards Women

Martina Galović, Damir Puškarić 63

Explicit Language in English Song Lyrics: Should We Be Worried?

Mila Bikić, Valerija Bočkaj 65

Comparison of the Representation of Male vs. Female Athletes in Croatian News Portals Using a NooJ Syntax Grammar

Klara Kozolić, Krešimir Štimac 67

Immigrant in the Light of Language Production

Barbara Vodanović 69

Semantic Analysis of Migrants' Self-Entrepreneurship Ecosystem Narratives

Cecilia Olivieri, Jie Sheng, Lorenzo Maggio Laquidara, Agathe Semlali ... 71

Reviewing the Position of the "Other" in Croatia's "Non-European" Collections in the Second Half of the 20th Century Using NLP

Martina Bobinac 74

Engaging with the Agenda 2030

Stella Nunzia Costanza, Antonio Pagano, Antonio Duca 76

NATURAL LANGUAGE PROCESSING APPLICATIONS 78

NooJ Grammars for Morphophonemic Continuity and Semantic Discontinuity Title

Mario Monteleone 79

Exploring Digital Literary Communication: Theoretical, Processual, and Environmental Perspectives with a Case Study of Linguistic Analysis and Graphical Representation of Dante Alighieri's Italian Language

Francesco Saverio Tortoriello, Ritamaria Bucciarelli, Ilaria Veronesi, Andrea Rodrigo 81

Machine Translation and Multi-Words Expressions: Pragmatic Analyzers Techniques Using NOOJ

Ali Boulaalam, Nisrine El Hannach 83

INVITED TALKS

From Corpora and Corpus Tools to Dictionaries and Beyond: Infrastructure for Slovene

Iztok Kosem

Centre for Language Resources and Technologies, University of Ljubljana &
Jožef Stefan Institute
Ljubljana, Slovenia
iztok.kosem@cjvt.si

Abstract

The Centre for Language Resources and Technologies at the University of Ljubljana has developed a wide range of resources and tools in the past decade, including corpora, dictionaries, syntactic parsers, text analyzers, and concordancers. These open-source resources and tools are an important part of language infrastructure for Slovene. The center also supports the research and wider community by providing APIs access to its resources, making data available in the CLARIN.SI repository, localizing external tools like NOOJ, compiling specialized corpora, and extracting datasets.

However, keeping up with the constantly and rapidly improving language technologies and methods is a major challenge for CJVT UL. Every new project introduces new approaches, techniques, and algorithms that must be integrated into the infrastructure. This requires a deep understanding of the latest developments in the field, as well as the human resources needed to implement and test these additions. Ensuring compatibility and interoperability with other language tools and platforms adds another layer of complexity.

To address this challenge, CJVT UL consolidated all its resources into a single digital database. This trend can be seen across Europe, including in Estonian (Tavast et al., 2018), German (Geyken, 2019), Polish (Żmigrodzki, 2018), and Dutch (Colman, 2016) language resources. Now, all the center's lexical resources, including a thesaurus, a collocations dictionary, and a bilingual Slovenian-Hungarian dictionary, draw from the same core set of lexical units and concepts. However, importing and integrating data from various lexicographic and linguistic activities, as well as from student projects and crowdsourcing, has been challenging.

In the presentation, the focus will be on data integration issues, such as merging data from different dictionaries and lexicons, working with corpora with different versions of annotation, and converting between different formats. The presentation will also cover recent projects, including the recent upgrade of several lexicographic resources, a monitor corpus, and automatic text categorization tools. The impact of ChatGPT's arrival on the center's work, particularly in the lexicographic context, will be discussed, and an experiment of its use in one of the center's projects will be presented.

Key words

Language infrastructure, Slovene, digital dictionary database, corpora, tools

References

[1] Colman, L. (2016) Sustainable Lexicography: Where to Go from Here with the ANW (Algemeen Nederlands Woordenboek, an online general language dictionary of contemporary Dutch)? *International Journal of Lexicography*, 29/2, 139–155.

[2] Geyken, A. (2019) The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography. In: Kosem, I., Zingano Kuhn, T. (eds.) *Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts*. Lexical Computing CZ s.r.o., Brno.

[3] Tavast, A., Langemets, M., Kallas, J., Koppel, K. (2018) Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In: Čibej, J., Gorjanc, V., Kosem, I., Krek, S. (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 749–761. Ljubljana: Znanstvena založba Filozofske fakultete.

[4] Żmigrodzki, P. (2018) Methodological Issues of the Compilation of the Polish Academy of Sciences Great Dictionary of Polish. In: Čibej, J., Gorjanc, V., Kosem, I., Krek, S. (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 209–219. Ljubljana: Znanstvena založba Filozofske fakultete.

Using Semantic Hypergraphs to Represent Knowledge in Natural Language Text

Branko Žitko

Faculty of Science, University of Split
Split, Croatia
branko.zitko@pmfst.hr

Abstract

The process of extracting knowledge from natural language text depends on the knowledge representation model. The goal of knowledge representation is to enable machines to reason while preserving the readability of knowledge for humans. Several knowledge representations have been used for this purpose, and in this speech, the focus will be on the semantic hypergraph (Menezes and Roth, 2019).

Using a semantic hypergraph for knowledge representation aims to preserve the complexity of natural language in a relatively simple formal structure. In this discourse, an in-depth description of the process of extracting knowledge from natural language text is presented. It begins with an introduction to the basic structural elements of the hypergraph and its expressive possibilities. It continues with the motivation for why the semantic hypergraph was chosen as a knowledge representation model, and compares it to related knowledge representations, such as semantic networks, predicate logic, and abstract meaning representation.

In our approach, the process of knowledge extraction consists of two phases. In the first phase, well-known natural language processing techniques are applied at the syntactic and semantic levels. We use sentence segmentation, tokenization, lemmatization, POS tagging, named entity recognition, dependency parsing, semantic role labeling, coreference resolution, predicate sense disambiguation (Žitko et al., 2022), and word sense disambiguation. Tools based mainly on machine learning are used to accomplish this phase. In the second phase, a parser that combines the obtained blocks of information from the previous phase into a semantic hypergraph is described. First, a semantic hypergraph annotation is determined for each word, and then the words are arranged into a hierarchical structure of the hypergraph.

In the final stages of discourse about knowledge extraction from text, the Natural Language to Semantic Hypergraph dataset (NL2SH) is presented. This dataset is used for the development and evaluation of our parser. NL2SH dataset contains 664 sentences labeled with natural language annotations, and for each sentence is assigned a semantic hypergraph. By analyzing errors, problematic language structures were identified, which will be used to tune our knowledge extraction process.

Finally, we will demonstrate our intentions to apply the semantic hypergraph for concept mapping, question generation and sentence simplification (Grubišić et al.,

2022). For this purpose, the original semantic hypergraph pattern language has been extended to enable searching and transforming the semantic hypergraph.

Key words

Semantic hypergraph, knowledge extraction, natural language processing

References

[1] Grubišić, A., Žitko, B., Gašpar, A., Vasić, D., Dodaj, A. (2022) Evaluation of Split-and-Rephrase Output of the Knowledge Extraction Tool in the Intelligent Tutoring System. *Expert Systems with Applications*, Vol. 187, Article 115875.

[2] Menezes, T., Roth, C. (2019) Semantic Hypergraphs. CoRR (abs/1908.10784).

[3] Žitko, B., Bročić, L., Gašpar, A., Grubišić, A., Vasić, D., Šarić-Grgić, I. (2022) Automatic Predicate Sense Disambiguation Using Syntactic and Semantic Features. In: *Proceedings of the Conference on Language Technologies and Digital Humanities* 2022, 227–234. Ljubljana, Slovenija.

MORPHOLOGICAL & LEXICAL RESOURCES

Quechua-Spanish and Spanish-Quechua Electronic Dictionaries of Verbs for NLP

Duran Maximiliano

Université de Franche-Comté
Besançon, France
duran_maximiliano@yahoo.fr

Abstract

The automatic processing of the Quechua language (APQL) needs an electronic dictionary of Spanish-Quechua-Spanish verbs. And yet, any bilingual NLP project within these languages requires this essential linguistic resource.

This article presents some advancements in the construction of such dictionaries.

My first challenge was to choose the Spanish dictionary to be used as the primary reference. Among the few Spanish monolingual digitalized dictionaries, I decided on the Spanish Module Argentine-Chile dictionary initially built at the Universidad Autónoma of Barcelona. It contains 6630 verbs and is entirely compatible with NooJ's formalism, and I decided to translate it into Quechua.

A significant difficulty is that the Quechua lexicon of simple verbs contains around 1,500 entries. How to match 6630 Spanish verbs with only 1,500 Quechua verbs?

In the present article, I show how I use the remarkable Quechua strategy of generating new verbs by suffix derivation, utilizing similar methods that I had applied in our previous work for the Quechua-French electronic dictionary at the University of Franche-Comté.

As a first step, I have inventoried all the Quechua suffixes and detailed their corresponding Spanish semantic values. This set of suffixes, which I call IPS_DRV, contains 27 elements. Thus, each Quechua verb, transitive or intransitive, gives rise to at least 27 derived verbs. Next, we need to formalize the paradigms and formal grammars that will allow us to obtain those derivations automatically. This was done with the help of the NooJ platform.

After parsing, these grammars generate 40,500 atomic linguistic units (CALU) that can be conjugated. I have used the printed dictionaries listed in the references and my introspection to translate manually many of the Quechua verbs that still need to be completed in the first stage. In the resulting electronic dictionary, I have included some relevant semantic tags to each Quechua verb and its corresponding inflection grammar.

On the other hand, I wrote an algorithm that allows the reciprocal translation from Spanish to Quechua of more than 6 000 derived verbs. So, I obtained the SP-QU electronic dictionary.

Key words

Quechua electronic dictionary, verbal suffixes, formalization of Qurchua verb, verb derivation, Quechua

References

[1] Duran, M. (2009) Dictionnaire Quechua-Castellano-Quechua. Editions HC, Paris.
[2] Duran, M. (2013) Formalizing Quechua Verbs Inflexion. In: Koeva, S., Mesfar, S., Silberztein, M. (eds.) *Formalizing Natural Languages with NooJ 2013*, 41–51. Cambridge Scholars Publishing.
[3] Duran, M. (2017) Di*ctionnaire électronique français-quechua des verbes pour le TAL*. Thèse doctorale. Université de Franche-Comté, Besançon.
[4] De Santo Tomas, D. (1560) *Lexicon, o Vocabulario de la lengua general del Perv.* Impresso en Valladolid por Francisco Fernandez de Cordoua.
[5] Gonçalez Holguin, D. (1608) *Vocabvlario de la lengva general de todo el Perv llamada lengua Qquichua o lengua del Inca*. Impresa en la Ciudad de los Reyes por Francisco del Canto.
[6] Guardia Mayorga, C. (1973) *Gramatica Kechwa*. Ediciones Los Andes, Lima, Peru.
[7] Perroud, P. C. (1970) *Diccionario castellano kechwa, kechwa castellano*. Dialecto de Ayacucho. Seminario San Alfonso, Santa Clara, Peru.
[8] Silberztein, M. (2010) La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. *Langages*, 179–180 (3–4), 221–241.

A NooJ Dictionary for Italian Light Verb Constructions

Alessia Nicola

Università degli Studi di Salerno
Salerno, Italy
alessia.giocondo@gmail.com

Giocondo Cirillo

Università degli Studi di Salerno
Salerno, Italy
nicirillo@unisa.it

Abstract

Light Verb Constructions (LVCs) are Multiword Expressions (MWE) occurring in many languages. LVCs have the canonical form *verb + noun* or *verb + adjective* and are characterized by the fact that the verbal component is semantically bleached while the noun/adjective conveys most of the meaning. Italian LVCs include *fare una foto, avere fame, dare una lavata, fare festa* etc. Just like collocations, LVCs are interpreted through lexical restriction. According to Bosque's compositional approach, there is a syntax-lexicon interface explaining the predicate-argument relationship: the predicates whether verbal, adjective, adverbial, or prepositional select their arguments. Unlike collocations, lexical restrictions in LVCs are activated by the noun that plays a predicative role. In summary, LVC can be defined as a structure where a verb combines with a noun predicate which provides the primary semantic content. Some constructions are more likely to be syntactically autonomous (e.g. rather than *fare una telefonata*, *fare festa* is less autonomous and has gone through lexicalization). According to Bratánková, the phenomenon could be explained by the loss of referentiality of the noun when it is not preceded by the article (Bratánková, 2013).

There has been an increasing interest in LVCs since this phenomenon occurs in many languages, and their identification is important also for natural language processing (NLP) tasks. Furthermore, LVCs exist alongside their synthetic verb counterparts, as long as *fare una foto* can actually be more efficiently replaced by its synthetic form *fotografare*. Some authors address the correspondence between LVCs and synthetic verbs from the NLP perspective (Chatzitheodorou, 2014; Mirto, 2021). For instance, Chatzitheodorou links LVCs to synthetic verbs, but without restructuring the sentence. Mirto, on the other hand, extracts the meaning from sentences with LVCs and synthetic verbs, but does not propose a technique to generate another sentence from that meaning representations. Our aim is to develop a tool that, given a sentence containing an LVC, generates a sentence with the syntetic verb. For example, given the sentence "Giovanni farà una festa alla sorella", the system produces the sentence "Giovanni festeggerà la sorella".

We also underline the existence of light verbs which are associated with a passive or reflexive synthetic form, for example, *provare delle emozioni* is more related to *emozionarsi* than to *emozionare*.

In this paper, we describe a NooJ dictionary containing the most common Italian LVCs. To develop it, we used the software NooJ. Firstly, we used Python to extract from the itWAC corpus (Baroni et al., 2009) all the sequences *verb + (preposition) + (determiner) + V-n* and *verb + V-a* (where *V-n* and *V-a* are nouns and adjectives morphologically related to verbs). Secondly, we computed a measure of the semantic

similarity between LVCs and their synthetic forms via a distributional semantic model. Sequences with frequency ≤ 100 and similarity ≤ 0.6 were filtered out. Finally, we manually deleted incorrect entries and associated each entry with the following properties: (1) its counterpart synthetic verb; (2) its similarity with the synthetic verb; (3) its frequency in the itWAC corpus; (4) the verb voice; (5) the lexical aspect; (6) the inflectional paradigm. We believe that this dictionary will foster the study of the LVC phenomenon and ease the automatic treatment of Italian LVCs with NooJ.

Key words

Light verb constructions, distributional semantic, NooJ dictionary

References

[1] Baroni, M., Bernardini, S., Ferraresi, A. et al. (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources & Evaluation* 43, 209–226.

[2] Bratánková, L. (2013) Le costruzioni italiane a verbo supporto. Un'analisi condotta sul corpus parallelo ceco-italiano. *Acta Universitatis Carolinae Philologica*, 2, 55–70.

[3] Chatzitheodorou, K. (2014) Paraphrasing of Italian Support Verb Constructions Based on Lexical and Grammatical Resources. In: *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, 1–7.

[4] Mirto, I. M. (2021) Natural Language Inference in Ordinary and Support Verb Constructions. In: *Distributed Computing and Artificial Intelligence, 17th International Conference*, 124-133. Springer International Publishing.

[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

Attenuative Collocations and Parametric Verbs in Old French and Old Spanish: A Contrastive Study

Xavier Blanco

Universitat Autònoma de Barcelona
Barcelona, Spain
Xavier.Blanco@uab.cat

Rafael García Pérez

Universidad Carlos III de Madrid
Madrid, Spain
rafael.garcia.perez@uc3m.es

Abstract

We have taken on a comprehensive description of the intensive collocations of Old French and Old Spanish as part of the Colindante project (see Blanco, 2020; Blanco and García Pérez, 2021). In addition, the attenuative (or anti-intensive) collocations must also be handled. A significant portion of these collocations include the pairing of a parametric verb with a noun that expresses little value. *Ne pas prendre an ail, Ne pas donner un bouton pour, Ne pas valoir une maille* are a few instances in French. Other instances in Spanish include: *no valer un higo, no valer una arveja, no haber/tener una meaja, no preciar un clavo*....

We will discuss the most prevalent expressive negation reinforcements found in electronic literary corpora of Old French and Old Spanish during our presentation. It is important to emphasize that these nouns were the direct objects of the parametric verb which became, through a process of grammaticalization, a part of a discontinuous negator embracing the verb. Thus, they cannot be considered the second semantic actant of the verb. Also, there is a collocational relationship between these nouns and the verb.

Regarding the linguistic nature of the aforementioned collocations, we will address a number of issues. One of these issues would focus on the metaphorical mechanisms that underlie the shift in meaning between the expression of a price and the expression of a minimum value or even of a zero value (pejoratively connoted).

By using the lexical-functional formalism, which was developed within the framework of Explanatory and combinatorial lexicology, we will propose a modelling approach to the aforementioned structures (see Mel'čuk and Polguère, 2021). Textual databases *Base textuelle Frantext* (1998–2022) and BFM (2019) will be used to extract the French lexical and syntactic data, while CORDE and CDH will be used to extract the Spanish data. Via the NooJ linguistic engineering platform, the lexical units and a portion of the corpus are implemented.

This communication proposal is part of the Colindante research project, *Ministerio de Ciencia e Innovación* (Spain).

Key words

Intensive collocations, anti-intensive collocations, old French, old Spanish, parametric verbs

References

[1] *Base textuelle Frantext* [en ligne]. (1998–2022) ATILF-CNRS & Université de Lorraine. Available at: https://www.frantext.fr/.

[2] Blanco, X. (2020) Remarques sur la variation diachronique des collocation. *Cahiers de Lexicologie*, 116, 71–94.

[3] Blanco, X., García Pérez, R. (2021) Las estructuras comparativas intensivas aplicadas al adjetivo *negro* en español medieval en comparación con el francés. *Romanica Olomucensia,* 33(1), 21–39.

[4] BFM – *Base de Français Médiéval* [en ligne]. (2019) ENS de Lyon, Laboratoire IHRIM, Lyon. <txm.bfm-corpus.org>.

[5] Mel'čuk, I., Polguère, A. (2021) Les fonctions lexicales dernier cri. In: Marengo, S. (éd.) *La Théorie Sens-Texte et ses applications. Lexicologie, lexicographie, terminologie, didactique des langues*, 75–155. L'Harmattan, Paris.

[6] *Banco de datos (CORDE)* [en línea]. *Corpus diacrónico del español.* Real Academia Española, Madrid. Available at: http://www.rae.es.

[7] *Corpus del Diccionario histórico de la lengua española (CDH)* [en línea]. (2013) Real Academia Española, Madrid. Available at: https://apps.rae.es/CNDHE.

[8] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

German *Selbst*-Compounds: A NooJ Grammar

Marco Angster

University of Zadar
Zadar, Croatia
mangster@unizd.hr

Abstract

German has been defined as a language "keen on compounding" (Gaeta and Schlücker, 2012). Compounds can be created combining a variety of word classes, including minor, closed ones, such as pronouns (see *Ichform* 'narration in first person', *Wirbewusstsein* 'group consciousness' from *ich* 'I' and *wir* 'we', respectively in Fleischer and Barz, 1995). Among pronouns Fleischer and Barz (1995) also list *selbst*, an element involved in heavy reflexive constructions (*Ich töte mich selbst* 'I kill mySELF'; cfr. its light counterpart *Ich töte mich* 'I kill myself'). König and Gast (2006) include however *selbst* in the cross-linguistically heterogeneous class of intensifiers, elements which focus on a participant (mostly a subject or object) and at the same time exclude the possible alternative referent which could perform or undergo the action.

Differently from other compounds formed with closed-class elements, compounds having *selbst* as modifier (e.g. *Selbsttötung* 'self-assassination', cfr. *Tötung* 'murder', from *töten*, 'to kill') have high type frequency and its family of complex words includes a variety of different bases, mainly deverbal: action nouns (*Selbstzerstörung* 'self-destruction'), agent nouns (*Selbstzerstörer* 'self-destroyer'), deverbal adjectives (*selbstzerstörlich* 'self-destructive'), participles in adjectival use (*selbstzerstörendes* Bild 'self-destructing picture', *ein selbstzerstörtes Leben* 'a self-destructed life'), etc., all linked to the verb *zerstören* 'destroy'.

It is usually considered that action nouns are the most type-frequent bases in selbst-compounds and that an interpretation by which the arguments of the underlying verb are coreferential is to be sought, so that these compounds have been labelled reflexive compounds (König, 2006). However, *selbst*-compounds can be modified by genitive constructions that constitute one of the participants of the underlying verb (*Selbsternte von Obst und Gemüse* 'self-harvest of fruit and vegetables', i.e. 'fruit and vegetables are harvested autonomously' and not 'fruit and vegetables harvest themselves').

The aim of this paper is twofold. Firstly, using NooJ, I will develop a grammar for *selbst*-compound which is able to classify the compounds extracted from a large corpus (the *deTenTen13* corpus of the *TenTen Corpus Family* in Jakubíček et al., 2013) according to their word class, derivational pattern, and underlying verb. Secondly, exploiting the grammar, I will compare the different rates of the various expected derivational patterns and of underlying verbs in different contexts, such as when *selbst*-compounds co-occur with genitive constructions, or they lack a syntactic modification.

Key words

Compounding, intensifiers, reflexives, corpus-based analysis

References

[1] Fleischer, W., Barz, I. (1995) *Wortbildung der deutschen Gegenwartsprache*. 2. durchgesehene und ergänzte Auflage. Niemeyer, Tübingen.

[2] Gaeta, L., Schlücker, B. (2012) *Das Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte*. De Gruyter, Berlin, Boston.

[3] König, E., Gast, V. (2006) Focused Assertions of Identity: A Tipology of Intensifiers. *Linguistic Typology*, 10, 223–276

[4] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V. (2013) The TenTen Corpus Family. In: *7th International Corpus Linguistics Conference CL*, 125-127.

[5] König, E. (2011) Reflexive Nominal Compounds. *Studies in Language*, 35(1), 112–127.

Spelling Error Detection and Correction for Arabic Using NooJ

Rafik Kassmi
Ibn Tofail University
Kénitra, Morocco
rafik.kassmi@gmail.com

Samir Mbarki
Ibn Tofail University
Kénitra, Morocco
mbarkisamir@hotmail.com

Abdelaziz Mouloudi
Ibn Tofail University
Kénitra, Morocco
mouloudi_aziz@hotmail.com

Abstract

Arabic is a highly structured and generative language where most words are derived from a root while following a pattern. It is also highly agglutinative and allows for a large number of affixes to be added to each word, thus increasing the number of possible words. This richness and complexity can be confusing and lead to the production of erroneous texts. These errors are generally divided into typographical, cognitive, or phonetic errors.

80% of the above errors are due to one or more of the following reasons (Damerau, 1964):

(1) Insertion error consists of adding extra character. For example, typing مكتتوب (makttūb) for مكتوب (maktūb, written), the letter ت (t) is additionally inserted.

(2) Deletion error resulting from the absence of a character. For example, typing مدسة (madsah) for مدرسة (madrasah, shcool), the letter ر (r) is missing.

(3) Substitution error consisting of replacing one character with another. For example, typing حديفة (ḥadifah) for حديقة (ḥadiqah, garden), the letter ق (q) is substituted by mistake by ف (f).

(4) Permutation error due to the exchange of characters. For example, typing برح (barḥ) for بحر (baḥr, see), the position of the letter ح (ḥ) is exchanged with the letter ر (r).

A spell checker is a tool that processes words to identify spelling errors and help correct them (Olani and Midekso, 2014). If it has doubts about the spelling of the word, it suggests possible alternatives. It can be interactive or automatic. The interactive spell checker detects misspelled words, proposes possible corrections for each of them and then allows the user to choose the correction. In contrast, the automatic spell checker automatically replaces the misspelled word with the most likely word without any interaction with the user. It is a standalone or integrated tool used to efficiently process natural language in many applications such as machine translators, OCR, search engines and word processors.

The aim of our research is to implement a spell checker for Arabic by taking advantages of the power of the NooJ platform and using its command line program noojapply. This spell checker will: first, detect all errors using the El-DicAr dictionary (Mesfar, 2006) combined with the improved morpho-syntactic grammar

handling agglutination (Kassmi et al., 2018) in NooJ. Then, generate corrections and candidate suggestions in NooJ. Next, rank the candidates in descending order.

Key words

Arabic language, spell checker, spelling errors, NooJ, El-DicAr

References

[1] Damerau, F. J. (1964) A Technique for Computer Detection and Correction of Spelling Errors. *Communication of the ACM*, 5(3), 171–176.

[2] Kassmi R., Mourchid M., Mouloudi A., Mbarki S. (2018) Processing Agglutination with a Morpho-Syntactic Graph in NooJ. In: Mbarki, S., Mourchid, M., Silberztein, M. (eds.) *Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications*, 40–51. Springer, Cham.

[3] Mesfar, S. (2006) Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. Phd Thesis. Franche-Comte University, Besançon.

[4] Olani, G., Midekso, D. (2014) Design and Implementation of Morphology Based Spell Checker. *International Journal of Scientific & Technology Research*, 3, 1–8.

[5] Silberztein, M. (2015) *La formalisation des langues : l'approche de NooJ*. ISTE Editions, London.

Towards a Linguistic Annotation of Arabic Legal Texts: A Multilingual Electronic Dictionary for Arabic

Khadija Ait ElFqih
L'Orientale University
Naples, Italy
Kaitelfqih@unior.it

Maria Pia di Buono
L'Orientale University
Naples, Italy
mpdibuono@unior.it

Johanna Monti
L'Orientale University
Naples, Italy
jmonti@unior.it

Abstract

Terminology translation plays a significant role in domain-specific machine translation (MT) (Štajner et al., 2016). However, some knowledge domains and languages still suffer from the lack of high-quality MT results, due to the mistranslation of terminology (Mediouni, 2016). This is the case in the legal domain and the Arabic language. Example 1 from the Moroccan family code shows the comparison between human translation (HT) and Google Translation (GT) taking the terms 'الفراش', 'الطعن', 'اللعان', 'القطع' into consideration:

> **AR:** يعتبر **الفراش** بشروطه حجة قاطعة على ثبوت النسب, لا يمكن **الطعن** فيه إلا من الزوج عن طريق **اللعان**, أو بواسطة خبرة تفيد **القطع**
>
> **GT:** The **Mattress**, with its conditions, is considered a definitive proof of Paternity, and it can only be **challenged** by the husband through **li'an**, or by means of experience that proves the **severance**.
>
> **HT:** Marriage **consummation** is considered a strong proof of paternity, it can be **rebutted** only by the husband through **accusation** or through a **certain** evidence.

The bold terms in example 1 are domain-specific and context-dependent, so their correct translation requires several aspects. Indeed, the HT has been produced considering the context, lexical, morphological, and semantic properties of the entries in addition to their equivalences across languages and legal systems. Whereas GT fails in its results.

This failure highlights the lack of terminology resources related to the legal domain (Zakraoui et al., 2020), the unfamiliarity of the legal systems to render the appropriate equivalences (Killman, 2014), and the terminology linguistic characteristics of this type of discourse (Varó and Hughes, 2014). This difficulty goes back to the need for more legal terminology resources. In fact, even though there are many Arabic legal dictionaries, most of them are not machine-readable and cannot be used in MT or other Natural Language Processing (NLP) applications.

Furthermore, there is a great need for terminology resources in which each entry is explicitly associated with a set of fully defined linguistic properties. For instance, NooJ is a free and strong linguistic software that gives the possibility of processing different natural languages and their related linguistic information including morpho-syntactic, and semantic information. In this paper, we present the development of a multilingual AR-EN-FR legal dictionary using NooJ that will be

capable of solving context-dependent issues, automatizing the process of annotating Arabic legal texts, and obtaining the automatic translation of technical legal terms into English and French. It contains 1080 single entries and compounds extracted from various legal documents mainly codes, decrees, constitutions, contracts, and provisions of different Arab countries.

As pipeline, we firstly extract our terms using NooJ grammars, then we proceed with the creation of our dictionary using NooJ morpho-syntactic information (part of speech, gender, number, etc.), syntactic information (transitive, intransitive, Naaqis), and the creation of our semantic tags that describe our domain-knowledge terms including legal, economy, Juri-religion, and geoUsage (following the ISO 20771:2020 standard for Legal translation Requirements) to indicate where a given term is adapted to express a legal practice. Finally, we propose the translations. In this phase, the process relies on consulting many portals including EUR-Lex, EuroVoc, and IATE to validate the equivalences of these terms among languages and legal systems. The formalization and compilation of our dictionary should enable the automatic annotation of any possible legal corpus in Arabic.

Key words

Legal terminology resources, multilingualism, Arabic legal dictionary, machine translation

References

[1] Killman, J. (2014) Vocabulary Accuracy of Statistical Machine Translation in the Legal Context. In: O'Brien., S., Simard., M., Specia., L. (eds.) 11th Conference of the Association for Machine Translation in the Americas, 85–98. Vancouver.

[2] Mediouni, M. (2016) Towards a Functional Approach to Arabic-English Legal Translation: The Role of Comparable/Parallel Texts. In: M. Taibi (ed.) New Insights into Arabic Translation and Interpreting, 115-160. Multilingual Matters, Bristol, Blue Ridge Summit.

[3] Štajner, S., Querido, A., Rendeiro, N., Rodrigues, JA., Branco, A. (2016) Use of Domain-Specific Language Resources in Machine Translation. In: Calzolari, N. et al. (eds.) The 10th International Conference on Language Resources and Evaluation (LREC'16), 592–598. Portorož.

[4] Varó, E. A., Hughes, B. (2014) Legal Translation Explained. Routledge, New York.

[5] Zakraoui, J., Saleh. M., Al-Maadeed, S., Mohamad AlJa'am, J. (2020) Evaluation of Arabic to English Machine Translation Systems. In: 11th International Conference on Information and Communication Systems (ICICS), 185–190. IEEE, Irbid, Jordan.

The Construction of a Multilingual Legal Ontology

Ismahane Kourtin

C.R.I.T. Laboratory
Bourgogne-Franche-Comté University
Besançon, France
kourtin_ismahane.math@yahoo.fr

Samir Mbarki

EDPAGS Laboratory, Faculty of Science
Ibn Tofail University
Kenitra, Morocco
mbarkisamir@hotmail.com

Abstract

The mass of information in the legal field, which is constantly increasing, has generated a capital need to organize and structure the content of the available documents, and thus transform them into an intelligent guide capable of providing complete and immediate answers to queries in natural language, and promoting the development of new forms of collective intelligence. Therefore, the question-answering system (QAS) (Hirschman and Gaizauskas, 2001), which is an application of the automatic language processing domain (NLP), perfectly meets this need by offering different mechanisms to provide adequate and precise answers to questions expressed in natural language. The general context of our work is the construction of a Question-Answering System in the legal field based on ontologies (Gruber, 1993; Borst, 1997), allowing users to ask a question on the desired information using natural language without having to browse through the documents. In this article, we will focus on the construction of a multilingual legal ontology based on legal laws and decrees (Mondary et al., 2008; Zaidi-Ayad, 2013). The legal ontology, which we propose to build from laws and decrees, will bring together the terminological material to optimize the automated management of laws and decrees, particularly during the stages of transforming users' questions in natural language into SPARQL queries on the one hand, and on the other hand when looking for answers to users' questions. We have adopted a methodological framework in seven steps for the construction of the legal ontology:

(1) Manual analysis of a sample of laws and decrees and the development of syntactic grammars for the extraction of the legal entities: in this step we build a sample of laws and decrees from which we manually extract the legal entities. Then we study the syntactic forms of the extracted legal entities, and we develop the syntactic grammars associated with NooJ which will be used to automatically extract legal entities from a corpus of laws and decrees.

(2) The constitution of a legal corpus: in this step a legal corpus is constituted from the legal laws and decrees.

(3) Linguistic analysis of the corpus and extraction of candidate legal entities: in this step we apply the syntactic grammars for the extraction of the legal entities, on the legal corpus, to extract the candidate legal entities for the construction of the legal ontology.

(4) Filtering of the candidate legal entities: then, we proceed to the filtering of all the extracted legal entities in order to eliminate the noise of the automatic extraction with NooJ, and to leave only the valid legal entities.

(5) Extraction of relations between legal entities: in this step, we identify the semantic relations between legal entities.

(6) Conceptualization: after having established the list of the legal entities, we will proceed to group these entities into semantic classes by establishing a list of ontology concepts.

(7) The construction of the legal ontology: finally, we will proceed with the structuring of the concepts into a terminological network, thus building our legal ontology.

Key words

Legal ontology, question-answering system (QAS), natural language processing (NLP), NooJ, legal field, legal decrees and laws

References

[1] Borst, W. N. (1997) Construction of Engineering Ontologies for Knowledge Sharing and Reuse. PhD Thesis. Universiteit Twente, Enschede. Available at: http://doc.utwente.nl/17864.

[2] Gruber, T. R. (1993) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 199–220. Available at: http://secs.ceas.uc.edu/~mazlack/ECE.716.Sp2011/Semantic.Web.Ontology.Papers/Gruber.93a.pdf.

[3] Hirschman, L., Gaizauskas, R. (2001) Natural Language Question Answering: The View from Here. *Natural Language Engineering*, 7(4), 275–300. Available at: http://dl.acm.org/citation.cfm?id=973891.

[4] Mondary, T., Després, S., Nazarenko, A., Szulman, S. (2008) Construction d'ontologies à partir de textes : la phase de conceptualisation. *19èmes Journées Francophones d'Ingénierie des Connaissances* (IC), 87–98.

[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

[6] Zaidi-Ayad, S. (2013) Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran). These. Universite Badji Mokhtar, Annaba.

Automatic Translation of Continuous and Fixed Arabic Frozen Expressions Using NooJ Platform

Asmaa Kourtin

Computer Science Research Laboratory
Faculty of Science Ibn Tofail University
Kenitra, Morocco
asmaa.kourtin@yahoo.fr

Samir Mbarki

EDPAGS Laboratory
Faculty of Science Ibn Tofail University
Kenitra, Morocco
mbarkisamir@hotmail.com

Abstract

The language lexicon is not only made up of single words but also frozen expressions. Therefore, we should not be limited to the study of the vocabulary and the analysis of the lexical meaning of a language to process it. The language treatment must include the study of the syntactic meaning, including the study of frozen or idiomatic expressions.

The frozen or idiomatic expressions have attracted the attention of several researchers in the last few years, leading to much research on different languages. The global meaning of these expressions is not deduced by joining the meanings of their components (Gross, 1993), so there are some problems in both processes of understanding and translating them.

The translating process of frozen expressions from one language into another is a real challenge for the translator given the linguistic and pragmatic specificities of the phenomenon that obliges a translator to have a good knowledge of both languages and cultures.

Therefore, the fact that all languages have specific cultures that are different. Besides, there are some differences in such factors as religion, geographical locations, different ideologies, and social classes of languages and societies that make the process of understanding and translating frozen expressions from one language into another very difficult (Shojaei, 2012).

In this work, we aim to create an automatic translator of the modern Arabic frozen expressions that are continuous and do not admit variations (see Kourtin et al., 2021), from the Arabic language into French and English, using the NooJ platform. For this reason, we will start by enriching our lexicon-grammar tables created in (Kourtin et al., 2019) by the French and English translations of each frozen expression. Then, we will transform these tables into dictionaries by using the generating NooJ dictionaries program, from lexicon-grammar tables created in NooJ, in the creation process of our translator. In the end, we will finish by testing the efficiency of this translator in texts and corpora.

Key words

Lexicon-grammar tables, frozen expressions, automatic translation, modern Arabic, NooJ platform

References

[1] Gross, M. (1993) Les phrases figées en français. *L'Information Grammaticale*, N. 59, 36–41.

[2] Kourtin, A., Amzali, A., Mourchid, M., Mouloudi, A., Mbarki, S. (2019) The Automatic Generation of NooJ Dictionaries from Lexicon-Grammar Tables. In: Fehri, H., Mesfar S., Silberztein M. (eds.) *Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications*, 65–76. Springer, Cham.

[3] Kourtin, A., Amzali, A., Mourchid, M., Mouloudi, A., Mbarki, S. (2021) Lexicon-Grammar Tables for Modern Arabic Frozen Expressions. In: Bigey, M., Richeton, A., Silberztein, M., Thomas, I. (eds.) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, 28–38. Springer, Cham.

[4] Shojaei, A. (2012) Translation of Idioms and Fixed Expressions: Strategies and Difficulties. *Theory and Practice in Language Studies*, 2(6), 1220–1229.

[5] Silberztein, M. (2015) *La formalisation des langues : l'approche de NooJ*. ISTE Editions, London.

Recognition of Frozen Expressions in Belarusian NooJ Module

Yauheniya Zianouka
Mikita Suprushuk
David Latyshevich
Yuras Hetsevich

United Institute of Informatics Problems
Minsk, the Republic of Belarus
ssrlab221@gmail.com

Abstract

In modern science, the question of the initial structural unit and perception of speech has no unambiguous solution because of various approaches and principles. However, it is known that speech has a syntagmatic nature and comprises lexical units that form syntagms. The author's delimitation and right intonation provide an adequate perception of speech. But synthesized speech, which is presented in various applications with voice accompaniment, is absorbed as unnatural, illegible, and inexpressive. The way to solve this problem is to develop specific methods and algorithms for analyzing and processing intonation features of natural speech, its automatic syntagmatic separation and implementation of all intonation constructions of a given language in NooJ. It will lead to automated reproduction of arbitrary text with the manner of human reading, and not an artificial system.

At the previous stages of our research, we have composed syntactic grammars for extracting syntagms at the punctuational and lexical levels, highlighting the intonation boundaries (Hetsevich et al., 2016; Zianouka et al., 2021). The next task is to form a syntactic grammar for delimiting phraseological units or so-called "Frozen expressions". These phenomena are reproducible, at least two-component linguistic units that combine with words of free use and are integral in meaning. As a rule, they are stable in their composition and structure (that is idioms). But in Belarusian, there are many frozen expressions that complicate their search due to structural units. For instance, the composition of phraseological expressions can be replaced by synonyms or other separate words. Or combinations, where one of the components is used in a phraseologically related meaning, and the other in a free one. Another problem is the order of units: in the phraseology, it can be fixed or, more often, it is used with the reverse order.

So, it is planned to compile a phraseological dictionary of the Belarusian language in NooJ format based on Etymological dictionary of Belarusian phraseological units and annotate it. The next step is to build syntactic grammars for searching the most typical groups of frozen expressions. And finally, to test them on the literary corpus of Belarusian NooJ module. This will contribute to the study of automatic processing of phraseological units for their further extraction into separate syntagms and forming their intonation portraits.

Acknowledgements. The research is prepared under the project №20213054 of The Belarusian Republican Foundation for Fundamental Research.

Key words

Frozen expression, syntagma, intonation, delimitation, automatic processing

References

[1] Hetsevich, Y., Okrut, T., Lobanov, B. (2016) Grammars for the Sentence into Phrase Segmentation: Punctuation Level. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.) *Automatic Processing of Natural-Language Electronic Texts with NooJ*, 74–82. Springer, Cham.

[2] NooJ: A Linguistic Development Environment, http://www.nooj-association.org, last accessed 2023/01/03.

[3] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

[4] Zianouka, Y., Hetsevich Y., Latyshevich, D., Dzenisiuk, D. (2021) Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module. In: Bigey, M., Richton, A., Silberztein, M., Thomas, I. (eds.) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, 231-242. Springer, Cham.

A Prototype of Indonesian Multi-Level Tagger: SANTI-Network

Prihantoro

Universitas Diponegoro
Kota Semarang, Indonesia
prihantoro@live.undip.ac.id

Abstract

In this paper, I present a prototype of the SANTI network. SANTI refers to *Sistem ANalisis Teks Indonesia* or Text Analysis System for Indonesian in English. The network is a multi-level tagger system for Indonesian, implemented using NooJ. The network is projected to integrate a morphological analyzer, POS tagger, semantic tagger, and syntactic parser of Indonesian. Today, two sub-systems are already available for use. The morphological analyzer for Indonesian, namely SANTI-morf (Prihantoro, 2021; Prihantoro, 2022) available for NooJ users, has been completed. SANTI-morf is the first sub-system. An Indonesian POS tagger is also available for NooJ users. This is the second sub-system. The creation of SANTI-sem, the semantic tagger for Indonesian whose annotation scheme adheres to the USAS tagset (Rayson, 2008), is in progress. This will be the third sub-system of the SANTI-network.

One of the core resources for the Indonesian semantic tagger, SANTI-sem, is a semantic dictionary of Indonesian, which now contains 4000+ entries ambiguously labeled. I here present two integration models and analyze how much they can be useful. The first one is to have three separate levels of tags for a word, namely morphology, morphosyntactic, and semantic. The second one is to have only two separate levels of tags: morphology and word. Tags for the word include both semantic and morphosyntactic. The main challenge of the two models is how to resolve ambiguities. My approach to addressing this issue is using the second model, which has a less annotation layer. The disambiguation is focused on the semantic level as the POS of words whose meanings are ambiguous are usually identical. Thus, morphosyntactic and semantic annotation resources shall apply first, followed by the application of morphological analyzer resources. Upon several simulations, this configuration gives the best result. While not without challenges, this method sheds light on integrating the next sub-system, SANTI-parse.

Keywords

Annotation, Indonesian, morphology, morphosyntactic, semantic

References

[1] Prihantoro. (2021) An Automatic Morphological Analysis System for Indonesian. Lancaster University, Lancaster.
[2] Prihantoro. (2022) SANTI-Morf Dictionaries. *Lexicography*, 9(2), 175–193.
[3] Rayson, R. (2008) From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics*, 13(4), 519–549.

[4] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.
[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

NooJ Dictionary for Rromani Language: Importing of a Published Dictionary to the NooJ System

Masako Watabe

C.R.I.T., Université de Franche-Comté
Besançon, France
masakowatabe@free.fr

Abstract

This paper aims at presenting the process of importing a published Rromani dictionary to the NooJ dictionary of the Rromani language.

The NooJ module for Rromani was created from the lexicon of two tiny corpus each of which is in a specific dialect. Then, the main inflectional paradigms including diasynonyms (dialectal variants) of nouns, verbs, adjectives, and some grammatical words such as personal pronouns have been programmed to formalize the morphology of this language. And now, we import a published dictionary of Rromani "Morri angluni rromane ćhibăqi evroputni lavustik" (Courthiade et al., 2009) to complete the NooJ dictionary of Rromani.

The Rromani module shows the innovative characteristic of being polylectal and thus including the four dialects of this language: named O-bi, O-mu, E-bi and E-mu. The division into four dialects is defined by two types of non-areal isoglosses crossed: 1) the opposition "o" vs. "e" which in a verbal ending (e.g. phirdom vs. phirdem 'I' walked), 2) a phonological mutation affecting two palato-alveolar affricates (e.g. ʒukel 'dog' [ʤukel] mutated into [ʒukel]). Each of these two dialect phenomena is associated with several other features and forms a bundle of isoglosses (i.e. diasystem).

The Rromani dialect system has a complex structure, yet its diasystem shows systematic correspondences of diasynonyms at the lexical, phonological and morphosyntactic level between the dialects. This fact encourages us to develop a single and common module for the entire Rromani language.

The published dictionary mentioned above is also polylectal. Lexical and morphological diasynonyms are included and their dialect properties are clearly indicated. Also, some phonetic variants are explained.

The problem we met is that no category is marked in this dictionary. Yet there are some signs indicating three categories: noun, adjective and verb. If a mark of gender (masculine or feminine) is inserted between singular and plural forms, this entry word is a noun. If the mark of plural "pl." is inserted between singular and plural forms, it is an adjective. If inflectional endings of present and past tense in the 3rd person singular (there is no infinitive in the Rromani) are indicated, it is a verb. Anyway, some grammatical knowledge is required to correctly recognize the categories. Concerning invariable words such as adverbs, conjunctions, appositions, it would be necessary to consult another dictionary to define their categories.

Another specificity of the NooJ module for Rromani is that we have created a double tag system to formalize the four dialects. If it is a word (or an inflectional form) used in two dialects that are defined by a common isogloss (e.g. O-bi and O-mu), this word (or form) will be annotated by a single tag (e.g. "rro"). If it is a word (or a form) used in a single dialect (e.g. O-bi) that is defined by two isoglosses, this word (or form) will be annotated by a double tag (e.g. "rro+rrbi"). Thus, we could formalize the Rromani diasystem and annotate correctly the diasynonyms in NooJ.

Not only dialect properties, but also diasynonyms are indicated for each of the concerned entry words in the NooJ dictionary of Rromani. Thus, on the one hand, learners of the Rromani language could understand more easily despite the difference in dialects. On the other hand, native speakers could discover the diasynonyms in various dialects and manage to communicate better with speakers of other dialects. That is important from a didactic point of view. In the future, we would like to produce didactic tools using this module as language resources to contribute not only to learners but also to native speakers of the Rromani language.

Key words

Rromani language, polylectal, diasystem, dictionary, NooJ

References

[1] Courthiade, M. et al. (2009) Morri angluni rromane ćhibăqi evroputni lavustik. Romano Kher, Budapest.
[2] Lava-laćhărne an-i khetani rromani ćhib. (2007) INALCO, Paris.
[3] R.E.D.-RROM (Restoring the European Dimension of Rromani Language and Culture) (2009-), www.red-rrom.com.
[4] Sarău, G. (2012) Dicţionar Rrom-Român. SIGMA, Bucarest.

Deciphering the Nomenclature of Chemical Compounds in NooJ

Kristina Kocijan
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
krkocijan@ffzg.hr

Krešimir Šojat
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
ksojat@ffzg.hr

Tomislav Portada
Ruđer Bošković Institute
Zagreb, Croatia
Tomislav.Portada@irb.hr

Abstract

Names of chemical compounds are found in the myriad of texts from different domains and thus pose an important language element that, due to its complexity, needs a layered approach within the NLP.

What we propose in this project is a multifaceted approach, the design of which is well supported by NooJ. The Croatian Chemical Compounds Module consists of three layers: it uses (1) the NooJ dictionary as the basis for (2) the morphological grammar, and both (1) and (2) are used for (3) the syntactic grammar.

We start with the dictionary of basic chemical elements (Kocijan et al., 2020), those found in the periodic table of elements, including the derivational grammar that helps us produce all the adjective forms derived from main nouns (e.g. *kalcij,* N -> *kalcijev,* A [calcium, N -> calcium, A]*; amonij,* N -> *amonijev,* A [ammonium, N -> ammonium, A]). Next, morphological grammar is designed to recognize single-unit words, homoatomic entities, denoting different variations of chemical names through a variety of suffixes (e.g. *sumpor,* N -> *sumpor**ast,*** A*, masculinum* | *sumpor**na,*** A*, femininum* [sulfur, N -> sulfurous, A |sulfuric, A]; *klor*, N -> *klor**id***, A | *klor**ovodičan,*** A [chlorine, N -> chloride, A | hydrochloride, A]), but also multiplicative prefixes (di-, tri-, tetra-, bis-, tris-, tetrakis-) used for simple and complicated entities (e.g. dioxygen, trichloride).

The algorithm for the detection of MWU, at this stage mainly binary compounds (Kocijan et al., 2021; Kocijan and Šojat, 2022), is designed within NooJ syntax grammar. This grammar uses both the main dictionary and the morphological grammar to recognize complex chemical compounds [*aluminijev oksid; kositrov karbonat; sumporna kiselina; željezov(II) hidrogensulfat, željezov(II) sulfat).* We are gradually developing the grammar to recognize even the most complex of compounds (Portada, Stilinović, 2007) that use digits, words, dashes, brackets, and commas (e.g. *(6E,13E)-18-brom-12-butil-11-klor-4,8-dietil-5-hidroksi-15-metoksitrikosa-6,13-dien-19-in-3,9-dion*).

Our main goal is to recognize a full compound, but also to segment it down to the smallest elements in order to produce its chemical formula. This information can be used in search queries but also in machine translations. The results are also valuable in the building of specialized lexica and systematization of scientific nomenclature.

Key words

Chemical compounds, nomenclature, morphology, syntactic grammar, NooJ

References

[1] Kocijan, K., Kurolt, S., Mijić, L. (2020) Building Croatian Medical Dictionary from Medical Corpus. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46(2), 765–782.

[2] Kocijan, K., Šojat, K. (2022) Formalizing the Recognition of Medical Domain Multiword Units. In: Dash, S., Parida, S., Tello, E., Acharya, B., Bojar, O. (eds.) *Natural Language Processing in Healthcare: A Special Focus on Low Resource Languages*, 89–120. CRC Press, Boca Raton.

[3] Kocijan, K., Šojat, K., Kurolt, S. (2020) Multiword Expressions in the Medical Domain: Who Carries the Domain-Specific Meaning. In: Bekavac, B., Kocijan, K., Silberztein, M., Šojat, K. (eds.) *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities,* 49–60. Springer, Cham.

[4] Portada, T., Stilinović, V. (2007) Što treba znati o hrvatskoj kemijskoj nomenklaturi? *Kemija u industriji*, 56(4), 209–215.

[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach.* Wiley-ISTE, London.

Croatian Cognition Verbs in Machine Sentence Processing

Marta Petrak
Faculty of Humanities and Social Sciences
Zagreb, Croatia
mpetrak@ffzg.hr

Bojana Mikelenić
Faculty of Humanities and Social Sciences
Zagreb, Croatia
bmikelen@ffzg.hr

Marko Orešković
National and University Library
Zagreb, Croatia
moreskovic@nsk.hr

Abstract

The paper presents a morphosemantic and syntactic analysis of Croatian cognition verbs formed from the roots *misl* 'think', *mozg* 'brain', *pamet* 'intelligence' and *um* 'mind'. Cognition verbs make a semantic category recognized in the literature as a whole whose semantic specificities influence its syntactic behaviour (e.g. Fellbaum, 1999; Hudeček, 2001; Grossman et al., 2002).

A list of verbs formed with the aforementioned roots and a minimum frequency of 10 occurrences was retrieved from hrWaC (Ljubešić and Klubička, 2014). This resulted in a total of 31 verbs. Most of these verbs were formed through prefixation (cf. Babić, 2002), such as for example *pomisliti* 'think of', *razmisliti* 'think through', *promozgati* 'ponder over', but there are also several examples of suffixation (e.g. *pametovati* 'lecture somebody') and compounding (*dvoumiti se* 'hesitate').

Based on the context in which they are used in the corpus, the argument structure of these verbs will be analysed. The verbs are mostly transitive or reflexive, but they present a lower degree of transitivity (Hopper and Thompson, 1980), due primarily to their abstractness. Other than the direct object, these verbs frequently accept different kinds of prepositional or non-prepositional indirect objects. We will analyse how the type of object correlates to the verb's prefix (if applicable) and aspect, and compare all the verbs in this regard.

The paper also brings a detailed analysis of the morphosemantic, i.e. derivational motivation (cf. Raffaelli, 2013) of these verbs. The semantic structure of the verbs demonstrates regularities in that they are frequently based on the conceptual metaphor mind is a container (for thoughts) and thoughts for mind / thoughts for cognitive processes conceptual metonymy.

Information obtained through this piece of research will be incorporated in the publicly available syntactic-semantic framework (SSF) (Orešković, 2019), which is continually enlarged with new elements. SSF already contains numerous grammatical properties (so-called T-structure), as well as many polysemous words from its 1.2 million token dictionary owing to its semantic domains created from publicly available online encyclopaedias. Special attention in this study is devoted to adding and expanding properties of Croatian verbs in SSF, especially those related to argument structure and semantic mechanisms which verbs were formed through. Verbal syntactico-semantic properties are of vital importance for achieving sentence

grammaticality, recognition of sentence structure and determining predicate relations.

Key words

Cognition verbs, machine sentence processing, morphosemantic analysis, syntactic analysis, Croatian

References

[1] Babić, S. (2002) *Tvorba riječi u hrvatskome književnome jeziku*. Globus, Zagreb.

[2] Fellbaum, C. (1999) The Organization of *Verbs and Verb Concepts in a Semantic Net.* In: Saint-Dizier, P. (ed.) *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, 98–110. Springer, Heidelberg.

[3] Grossman, M., Koenig, P., DeVita, C., Glosser, G., Alsop, D., Detre, J., Gee, J. (2002) Neural Representation of Verb Meaning: an fMRI Study. *Human Brain Mapping* 15(2), 124–34

[4] Hopper, P. J., Thomson, S. A. (1980) Transitivity in Grammar and Discourse. *Language* 56(2), 251–299.

[5] Hudeček, L. (2001) Glagoli govorenja i mišljenja u hrvatskome čakavskom književnom jeziku do 17. stoljeća – strani sintaktički utjecaji. *Rasprave IHJ*, 27, 95–112.

[6] Ljubešić, N., Klubička, F. (2014) {bs,hr,sr}WaC – Web Vorpora of Bosnian, Croatian and Serbian. In: *Proceedings of the 9th Web as Corpus Workshop* (WaC-9), 29–35. ACL, Gothenburg.

[7] Orešković, M. (2019) An Online Syntactic and Semantic Framework for Lexical Relations Extraction Using Natural Language Deterministic Model. PhD thesis. Faculty of Organisation and Informatics, Varaždin.

[8] Raffaelli, I. (2013) The Model of Morphosemantic Patterns in the Description of Lexical Architecture. *Lingue e linguaggio* 12(1), 47–72.

NooJ Dictionary of Croatian and English Internet Slang

Ivan Cota

Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
icota@m.ffzg.hr

Abstract

How often do you catch yourself googling the meaning of a term that you have just read somewhere on the Internet? A lot of the time we humans cannot understand phrases and abbreviations we read online without clarification, so it is no surprise our linguistic tools cannot understand them either. The goal of this project was to build a resource for NooJ, specifically a dictionary that would help recognize and label Internet slang terms. It came to our attention that the main source of all newly built corpora is based on Internet texts which is now more than ever littered with new Internet slang terms that cannot be overlooked because they often carry most of the meaning of a sentence or can even be the entire sentence.

Corpus used for development of the dictionary was "scraped" using python code and a Reddit API, on the Croatian subreddit, using 77 of the most popular threads that were chosen to best represent the Croatian Internet language and terminology. After the scraping of the corpus, the text was manually edited, mostly removing emojis, special characters, empty lines etc., things that would not be useful for building a dictionary. The final version of the corpus has three sub-corpora divided according to the year of posting starting with 2020 and ending with the latest postings in 2022, with the 2020 corpus having 39 153 tokens, the 2021 corpus having 94 309 and the 2022 corpus having 234 351 tokens.

Because of the bilingual nature of online communication, a significant number of words were standard English, so all the available tools for Croatian and English languages were applied to the text in the linguistic analysis. After the analysis, NooJ flagged 14 661 tokens as UNKNOWN. This list of words was then ordered by the frequency of appearance in the corpus and was cut down several times manually to provide only relevant (more than 2 appearances) cases of what authors decided was a slang term. This provided a list of 99 words that became the NooJ Internet slang dictionary (slang.dic) which can recognize 987 different word forms. Slang words in this list were then described based on: POS (noun, verb etc.), slang type ("razgsl" for common slang or "intsl" Internet slang), language ("en" for English, "hr" for Croatian) and flexional paradigm. There are 64 Croatian and 35 English slang terms. Of the total number of words there are 52 slang abbreviations in the dictionary, of which 17 are Croatian and all the English terms are abbreviations.

Due to difficulties in differentiating common from internet slang, the words are tagged according to the decision of the authors. The need for all of these classification categories came from the bilingual nature of the corpus and the recognized difference in slang form, particularly because there is a distinct difference between abbreviated slang expressions (e.g. "imo", "tbh") and, for

instance, everyday Croatian slang (e.g. “fakat”, “profa”). If expanded upon, the dictionary could be a useful tool for future linguistic analysis in research based on internet text corpora.

Key words

NooJ, slang, Internet, Reddit, dictionary

References

[1] Aoughlis, F., Métais, E. (2007). A Computer Science Electronic Dictionary for NooJ. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (eds.) *Natural Language Processing and Information Systems*. NLDB 2007. Lecture Notes in Computer Science, Vol 4592, 458–462. Springer, Berlin, Heidelberg.

[2] Manuel, K., Indukuri, K. V., Krishna, P. R. (2010) Analyzing Internet Slang for Sentiment Mining. In: *Second Vaagdevi International Conference on Information Technology for Real World Problems*, 9–11. Warangal, India.

[3] Monteleone, M. (2019) NooJ Grammars and Ethical Algorithms: Tackling On-Line Hate Speech. In: I. M. Mirto, M. Monteleone, & M. Silberztein (eds.) *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*, 180–191. Springer International Publishing.

[4] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

[5] Vučković, K., Tadić, M., Bekavac, B. (2010) Croatian Language Resources for NooJ. *Journal of Computing and Information Technology*, 18(4), 295–301.

Latin Pronouns, Numbers and Prepositions in the NooJ Tool

Anita Bartulović

University of Zadar
Zadar, Croatia
abartulo@unizd.hr

Linda Mijić

University of Zadar
Zadar, Croatia
lmijic@unizd.hr

Abstract

This paper contributes to the work on the building of a NooJ module for the Medieval Latin language. After creating NooJ resources for nouns, adjectives, and adverbs (see Mijić, Bartulović, 2020; two papers in the peer review process), the emphasis is now placed on the rest of the nominal forms (pronouns and numbers) as well as on prepositions. The paper presents dictionaries and inflectional grammars for recognizing inflectional forms of pronouns (personal, possessive, reflexive, demonstrative, relative, interrogative, indefinite) and numbers, as well as morphological grammars for recognizing complex (indefinite) pronouns (*aliqui*, *quidam*, etc.), and derivational grammars for different types of numerals (cardinal, ordinal, distributive, adverbial). After expanding the dictionary also with prepositions, we created syntactic grammar for the rection of prepositions, i.e. extraction of prepositional phrases which can include a various number of different types of inflectional words. Particular attention was paid to the medieval local peculiarities of the Latin language (see Stotz, 1996–2004). Compiled grammars are applied to a smaller corpus of the last wills and testaments written in Zadar commune in Medieval Latin. The results show a very high level of recognition, and these resources can be added to a future module for the Latin language.

Furthermore, the paper researches the use of adjectives and some pronouns as attributes in legal discourse, considering the formulaic nature of notary records which are not inclined to use a high range of attributes (especially adjectives). Namely, when conducting medieval last will and testament, the testator sometimes used an indefinite pronoun (*quidam*, *aliqui*) with the name of a person. The results of collocation analysis confirm that in some cases the indefinite pronoun is used in the transferred meaning ("some" = "illegal") and some rarely used adjectives (e.g. *dilectus*) show the emotional affection of some testators in higher degree towards certain people as recipients of their bequests.

Key words

Medieval Latin last wills and testaments, morphological grammars, the Latin language, NooJ

References

[1] Mijić, L., Bartulović, A. (2021) Formalizing Latin: An Example of Medieval Latin Wills. In: Bekavac, B., Kocijan, K., Silberztein, M., Šojat, K. (eds.)

Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities, 24–36. Springer, Cham.
[2] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.
[3] Stotz, P. (1996–2004) *Handbuch zur lateinischen Sprache des Mittelalters*, Vol. 1–5. Verlag C. H. Beck, München.

SYNTACTIC & SEMANTIC RESOURCES

Disambiguation Grammars for the Ukrainian Module

Olena Saint-Joanis

C.R.I.T., Université de Franche-Comté
Besançon, France
olena.saintjoanis@gmail.com

Abstract

Ukrainian is a flexional language. It means that the endings of lexical units change depending on their grammatical class (verb, noun, pronoun, adjective, etc.) as well as depending on their role in the sentence. Flexions are described by the flexion paradigms. However, some paradigms have identical endings for different cases. For example, in the nominal paradigm - "СИН *[SYN] sun*" - the Genitive and Accusative lexical units have the same ending - "a", while the Dative, Locative and Vocative units have the same ending "y":

(1) <E>a/Genitive+Singular| <E>a/Accusative+Singular|
(2) <E>y/Dative+Singular|<E>y/Locative+Singular|<E>y/Vocative+Singular|

This phenomenon is very common and can even be complicated in some cases, for example for adjectival paradigms, where the category of animate/inanimate objects multiplies duplicate endings: for masculine adjectives Genitive animate object = Accusative animate object, when Accusative inanimate object = Nominative inanimate object.

To solve this problem, it is necessary to build syntactic grammars that are capable of making a difference. Thus, we construct some grammars capable of case disambiguation and a separate grammar for recognizing gender in the noun group. We also offer a grammar for the verbal group and another for indeterminate pronouns.

We get very good results even for groups consisting of several lexical units. We are also thinking about expanding our work to cover the sentences, formed with a single variable lexical unit.

Key words

Ukrainian, disambiguation, syntactic grammars

References

[1] Gorpynych, V. O. (2004) *Morphologiya ukraïnskoï movy.* Akademiya, Kyïv.
[2] Plushch, M. Y. (2010) *Gramatyka ukraïnskoï movy. Chastyna 1. Morfemika. Slovotvir. Morfologiya. Pidruchnyk dlya studentiv filologichnyhk spetsialnostei vyshchyyh navchalnyh zakladiv.* Vyshcha shkola, Kyïv.
[3] Silberztein, M. (2015) Joe loves Lea: Transformational Analysis of Transitive Sentences. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.)

Automatic Processing of Natural-Language Electronic Texts with NooJ. Springer, Cham.
[4] Vyhovanets, I. R., Gorodenska, K. G. (2004) *Teorretychna morfologiya ukraïnskoï movy*. Pulsray, Kyïv.

A Proposal for the Processing of the Nucleus Verb Phrase of Pronominal (SVNPr) Verbs in Spanish

Andrea Rodrigo
CETEHIPL, Facultad de Humanidades y Artes, UNR
Rosario, Argentina
andreafrodrigo@yahoo.com.ar

Rodolfo Bonino
CETEHIPL, IES N°28 "Olga Cossettini"
Rosario, Argentina
rodolfobonino@yahoo.com.ar

Silvia Reyes
CETEHIPL, Facultad de Humanidades y Artes, UNR
Rosario, Argentina
sisureyes@gmail.com

Abstract

The *Centro de Estudios de Tecnología Educativa y Herramientas Informáticas de Procesamiento del Lenguaje* (CETEHIPL) focuses on the pedagogical application of the NooJ tool created by Silberztein [5]. The central ideas of this proposal are developed in our book *Aprendo con NooJ* [4]. In this line of work and always based on our Spanish module Argentina available on the NooJ platform (http://www.nooj4nlp.org/resources.html), we have been addressing the formalization of lexical categories such as the adjective and the adverb [2,3]. We will now go further into the processing of the verb, a complex lexical category that is a central element of the sentence. And when dealing with the verb in Spanish, it is essential to include clitics, since they always and only occur in construction with it. In this presentation, we will focus on the pronominal verbs of a specific corpus. However, since the complex nature of the verb in Spanish brings about that some verbs behave differently depending on their syntactic context, we will design dictionaries and grammars to account for this syntactic behavior.

Clitics may be defined as unstressed pronouns that depend phonologically upon a verb, are immediately adjacent to it and semantically related to its arguments. Clitics in Spanish show orthographic particularities that depend on verb conjugation and predicate polarity. Enclitics attach as inflections to and follow non-finite or non-personal verb forms (infinitives, gerunds, and participles) and positive imperative, forming a unit with them: infinitive *jactar**se*** (to boast), imperative second person *dé**se** por vencido* (*usted*) (give up). On the other hand, proclitics are orthographically separated, appear in front of and precede all other tensed finite forms: present indicative third person singular ***se** niega* (she/he/it refuses), *no **se** niega a (*she/he/it does not refuse to).

Many verbs are pronominal in some contexts and transitive in other contexts, such as the verb *dar* (to give) in the following example of our corpus: *Dame tu amor* (transitive verb) (Give me your love). But the same verb becomes pronominal (*darse*) when we say: *Me doy a conocer (*pronominal verb) (I make myself known). Following Bès [1], who describes the nucleus verb phrase in French, we will try to formalize the nucleus verb phrase of pronominal (SVNPr) verbs in Spanish.

According to our usual methodology, we will create a corpus of real texts written in Rioplatense Spanish. And specifically in this case, a corpus of popular songs that will allow us to account for the current use of pronominal verbs, as well as their idiosyncratic particularities. For the sake of formalization, we will introduce changes

in our grammars and dictionaries of the Spanish Module Argentina, and will develop a syntactic grammar to recognize and generate chains of clitics and verbs in the canonical word order of Spanish, preceded or not by negation. These latter strings as well as the study of verbs having other clitic combinations will be addressed in future work.

Key words

Clitics, pronominal verbs, NooJ, Spanish module Argentina, nucleus verb phrase

References

[1] Bès, G. G. (1999) La phrase verbal noyau en français écrit. Recherches sur le français parlé, vol. 15, 273–358.

[2] Rodrigo, A., Monteleone, M., Reyes, S. (2018) A Pedagogical Application of NooJ in Language Teaching: The Adjective in Spanish and Italian. In Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, 47–56. Santa Fe, New Mexico, USA. Available at: http://aclweb.org/anthology/W18-3807.

[3] Rodrigo, A., Reyes, S., Bonino, R. (2018) Some Aspects Concerning the Automatic Treatment of Adjectives and Adverbs in Spanish: A Pedagogical Application of the NooJ Platform. In: Mbarki, S., Mourchid, M., Silberztein, M. (eds.) Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications, 130–140. Springer, Cham.

[4] Rodrigo, A., Bonino, R. (2019) Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua. 1st edn. Editorial Ciudad Gótica, Rosario.

[5] Silberztein, M. (2016) Formalizing Natural Languages: The NooJ Approach. Wiley-ISTE, London.

A Rioplatense Spanish Date Grammar Using the NooJ Platform

Mariana González

CETEHIPL, Facultad de Humanidades y Artes, UNR
Rosario, Argentina
marianagonzalez826@gmail.com

Abstract

Since our project on using computer tools for pedagogical purposes in teaching and learning language (see Rodrigo and Bonino, 2019), our efforts are focused on finishing the dictionaries, the inflectional and the syntactic grammars in our Spanish-Argentina module so that language learners can make a metalinguistic reflection on the target language to generate linguistic knowledge in their interaction with a computer tool such as NooJ. With NooJ, a linguistic development software created by Max Silberztein, we work towards developing a grammar of dates which can automatically analyze everyday expressions describing place, date, and time in Rioplatense Spanish, as in the following examples:

(1) *en Rosario, a las 19 horas se estrenará la película más esperada (in Rosario, at 7 pm, there will be premiered the most awaited film)*
(2) *el martes pasado por la mañana llovió torrencialmente en la ciudad de Rosario (last Tuesday morning, there was a heavy rain in Rosario)*
(3) *el pasado miércoles a las 21 horas se produjo un robo en el centro (last Wednesday at 9 pm, there was an armed robbery downtown)*
(4) *a la hora de la siesta de ayer hubo un alerta meteorológico (yesterday during naptime, there was a weather alert)*

We noticed that the Argentinian Spanish module lacked a proper grammar to recognize common date expressions typical from Argentina and other Spanish speaking countries, we did not notice a significant difference as regards date expressions in Spanish variations. We took inspiration in a simple grammar that already existed in the Spanish module Argentina and inspired by the English date and time grammar available in the English module. We will validate such grammar using analysis and generation as two possibilities presented by NooJ. This will allow us to make any necessary corrections and adjustments to achieve satisfactory results, both the analysis and generation of grammatically well-formed statements in Rioplatense Spanish.

To conclude, we will propose a series of exercises in which language learners will use NooJ to examine expressions such as the ones above, guiding them using questions such as *at what time during the day do people have lunch, which expressions do we use to refer to the night or how do we refer to the time.* At the same time, we will challenge language learners to think in which possible way we can formulate these questions in syntactic grammar in NooJ to be analyzed or generated automatically. The questions will therefore drive learners to make a reflection on the language, as Lidia Usó Viciedo rightly states:

The current need to include, in the didactic field of first languages, both the metalinguistic reflection upon language and the inductive processes in teaching grammar providing and promoting its acquisition in the classroom. (Usó Viciedo, 2014: 49; the translation is ours).

Key words

NooJ, Spanish date grammar, Spanish module Argentina, pedagogical application, metalinguistic reflection

References

[1] Rodrigo, A., Bonino, R. (2019) *Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua*. 1st edn. Editorial Ciudad Gótica, Rosario.

[2] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

[3] Spanish Module Argentina. Available at: http://www.nooj4nlp.org/resources.html.

[4] Usó Viciedo, L. (2014) De la enseñanza tradicional de la gramática a la reflexión metalingüística en primeras lenguas. *Tejuelo. Didáctica de la Lengua y la Literatura. Educación*, Nº Extra 10, 49–64. Gobierno de Extremadura. Available at: https://dehesa.unex.es/handle/10662/8868.

Parafrasário: A Variety-Based Paraphrasary for Portuguese

Anabela Barreiro
INESC-ID
Lisboa, Portugal
anabela.barreiro@inesc-id.pt

Ida Rebelo
Universidad de Valladolid
Valladolid, Spain
ntrebelo@yahoo.com.br

Cristina Mota
INESC-ID
Lisboa, Portugal
cristina.mota@inesc-id.pt

Abstract

This paper introduces Parafrasário (paraphraser), a dataset of paraphrastic units at the multiword unit (see Barreiro, 2010), expression, and phrasal levels which is the multiword, expression, and phrasal equivalent to a standard dictionary, with canonical forms as its entries.

We start by describing a version of Parafrasário that combines paraphrastic units used in Portuguese from Portugal (PP) and Portuguese from Brazil (PB), aligned from PP–PB literary parallel corpora (Barreiro and Mota, 2017). The larger aim of Parafrasário is to integrate paraphrases from any variety of the Portuguese language. We discuss 5 different groups of contrast between paraphrases in terms of morphosyntactic and lexical choice, word order, and semantic similarity. Then, we present the methodology used to develop a first version of Parafrasário, including the identification of the Portuguese variety in the cases where multiwords, expressions or phrases are of specific usage in one of the varieties. This version of the Portuguese paraphraser contains approximately 1,500 entries and is available on the Multi3Generation COST Action (CA18231) website to be freely used by the research community.

Most entries (941) of the paraphraser are verbal construction of which (1) 23 entries are verbal on PP, but non-verbal on PB (e.g. *a transbordar de alegria* / *doido de felicidade* 'brimming over with happiness'), (2) 34 are verbal on PB, but non-verbal on PP (e.g. *no fim* / *quando terminam* 'when they finished'), and (3) all other entries (884) correspond to verbal constructions on both sides, PP and PB (e.g. *quando dá por isso* / *quando percebe* 'when she twigs it'). The remaining entries of the paraphraser are nominal/adjectival, adverbial, numeric expressions, among others.

In the rest of this paper, we focus on how to formalize in NooJ the paraphraser of verbal entries, which are more interesting and more productive for the purpose of creating transformational local grammars. Additionally, we show how to use this paraphraser in language generation in eSPERTo (see Mota et al., 2016), a NooJ based system for paraphrasing.

Key words

Paraphrases, paraphrasary, semantic equivalence, Portuguese NLP, language generation

References

[1] Barreiro, A. (2010) Make It Simple with Paraphrases. Automated Paraphrases for Authoring Aids and Machine Translation. Lambert Academic Publishing, Saarbrücken. Available at:
[2] https://www.linguateca.pt/Repositorio/AB-Thesis_030409.pdf
[3] Barreiro A., Mota C. (2017) e-PACT: eSPERTo Paraphrase Aligned Corpus of EN-EP/BP Translations. *Tradução em Revista*, 22, 87–102.
[4] eSPERTo: System for Paraphrasing in Smart Text Editing and Revision. Available at: https://www.inesc-id.pt/projects/IP02043/, last accessed 22/03/2023.
[5] Mota, C., Barreiro, A., Raposo, F., Ribeiro, R., Curto, S., Coheur, L. (2016) eSPERTo's Paraphrastic Knowledge Applied to Question-Answering and Summarization. In: Barone, L., Monteleone, M., Silberztein, M. (eds.) *Automatic Processing of Natural-Language Electronic Texts with NooJ*, 208–220. Springer, Cham.
[6] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

CORPUS LINGUISTICS & DISCOURSE ANALYSIS

The Limitations of Training Corpus-Based Methods in NLP

Max Silberztein

Université de Franche-Comté
Besançon, France
max.silberztein@univ-fcomte.fr

Abstract

Nowadays, most Natural Language Processing software applications use stochastic "black box" methods associated with training corpora to analyze texts written in natural languages. The stochastic approach has been so successful that it is universally seen as the only correct approach to processing natural languages, at the expense of linguistic methods: many linguistic centers have abandoned the scientific goal of formalizing natural languages with handcrafted electronic dictionaries and grammars; international scientific institutions equalize the notion of "linguistic resource" to "training corpus", and almost all papers presented at leading conferences such as COLING ("COmputational LINGuistics") or ACL (for "Association for Computational Linguistics") present systems or methods that contain no linguistic resources nor methods.

The principle at the basis of all stochastic approaches is the following: to analyze a sequence of words in a text, stochastic software looks for similar sequences in a training corpus, select among them the most "similar" one using some probabilistic, statistical, or neuron-network-based optimization, and then bring forth its analysis as the new sequence's analysis.

This paper argues that this principle (similar to using a cheat sheet in an exam without understanding the exam's questions) contains several flaws that will eventually stop stochastic software applications from gaining any progress. In particular, the paper shows that:

(1) so-called "reference" training corpora, such as the PennTreebank (see Taylor et al., 2003) and the COCA (see Kupść and Abeillé, 2008) contain many mistakes (see Dickinson, 2015; Silberztein, 2018; Volokh and Neumann, 2011).

(2) the poor tag sets used by stochastic taggers does not compare with the information provided by dictionaries used in NLP (see Kupść and Abeillé, 2008) and thus does not satisfy the most basic needs of NLP software applications such as Information Retrieval and Machine Translation.

(3) the very notion of "context" used by taggers to solve ambiguities automatically does not have any linguistic value (see Bontcheva et al., 2003; Brill and Wu, 1998; Schmid, 1994).

(4) the units processed by stochastic software are the wrong ones.

The paper shows how these flaws are translated into unreliable NLP software applications, and how using carefully handcrafted linguistic methods and resources could correct enhance them.

We finally dispute the scientific validity of the stochastic approach.

Key words

Linguistics, computational linguistics, corpus linguistics, natural language processing, NooJ

References

[1] Bontcheva, K., Maynard, D., Tablan, V., Cunningham, H. (2003) GATE: A Unicode-Based Infrastructure Supporting Multilingual Information Extraction. Available at: https://gate.ac.uk/sale/iesl03/iesl03.pdf.

[2] Brill, E., Wu, J. (1998) Classifier Combination for Improved Lexical Disambiguation. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol. 1, 191–195, Montreal, Quebec.

[3] Dickinson, M. (2015) Detection of Annotation Errors in Corpora. *Language & Linguistics Compass*, Vol. 9, Issue 3. Wiley Online Library. Available at: https://doi.org/10.1111/lnc3.12129.

[4] Kupść, A., Abeillé, A. (2008) Growing Treelex. In: *International Conference on Intelligent Text Processing and Computational Linguistics*, 28–39. Springer, Berlin, Heidelberg.

[5] Schmid, H. (1994) TreeTagger-a Language Independent Part-of-Speech Tagger. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.

[6] Silberztein, M. (2018) Using Linguistic Resources to Evaluate the Quality of Annotated Corpora. In: *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, 2–11.

[7] Taylor, A., Mitchell, M., Santorini, B. (2003) *The Penn Treebank: an Overview.* In: Abeillé, A. (ed.) *Treebanks. Text, Speech and Language Technology*, Vol 20, 5–22. Springer, Dordrecht.

[8] Volokh, A., Neumann, G. (2011) Automatic Detection and Correction of Errors in Dependency Treebanks. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technology*, 346–350.

Syntactic-Semantic Analysis of Perception Verbs in the Croatian Language

Daša Farkaš

Faculty of Humanities and Social Sciences
Zagreb, Croatia
dberovic@ffzg.hr

Kristina Kocijan

Faculty of Humanities and Social Sciences
Zagreb, Croatia
krkocijan@ffzg.hr

Abstract

Perception verbs express the internal ways and processes with which we perceive the world around us but also gather information from our surroundings. Their main feature is that they are ambiguous: they have their prototypical, physical meanings and non-prototypical, extended physical and metaphorical meanings. There are several studies of perceptive verbs for the Croatian language (Burić 2021; Mihaljević 2009, 2011; Raffaelli 2017), but none of them is corpus-based research that considers a larger number of verb lemmas that express perception verbs.

This research presents a corpus approach to verbs of perception for the Croatian language, which analyses them at the syntactic and semantic level with the help of NooJ. The project started by manual extraction of all verb lemmas related to perception verbs from the total list of verb lemmas from the Croatian Morphological Lexicon (Tadić and Fulgosi 2003). A total of 86 verbs were selected and divided into five semantic subgroups: *sight, hearing, taste, smell,* and *touch*. This information is marked for each NooJ dictionary entry adding the semantic tag *+prcp* to mark the semantic category of a **perception** verb, and its semantic subgroup *+viz, + sluh, +okus*, +miris, and *+dodir.* The verbs were next explored within three different domains (the corpus of medical texts, the corpus of parliamentary texts and the corpus of children's literature) to learn more about their syntactic and semantic features.

Regarding syntactic processing, the predicate complements of perception verbs were analyzed regarding the specific corpus and by individual groups of verbs. The assumption is that complements differ depending on whether the verb expresses a prototypical physical meaning or an extended metaphorical meaning. The ways in which predicate complements can be expressed will be shown in the paper.

The semantic analysis shows the distribution of verbs by specialized corpora and by the meaning of perception verbs. This analysis rests on the assumption that the distribution of prototypical physical meanings and extended metaphorical meanings differs depending on the specialized corpus in which it is observed, for example, that in the medical corpus there are more physical prototypical meanings, while in the corpus of children's literature there are more extended metaphorical meanings of perception verbs. Verbs are further classified according to how much they extend their meaning: from those that have only a prototypical physical meaning to those that develop more metaphorical meanings.

Key words

Perception verbs, Croatian language, predicate complements, prototypical physical meanings, metaphorical meanings

References

[1] Burić, H. (2021) Kognitivnolingvistički pristup sintaktičko-semantičkomu opisu osjetilnih glagola u hrvatskome jeziku. PhD thesis. University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb.

[2] Mihaljević, M. (2009) The Structure of Complements of Verbs of Perception in Croatian. In: Franks, S., Chidambaram, V., Joseph, B. (eds.) *A Linguist's Linguist: Studies in South Slavic Linguistics in Honor of E. Wayles Browne*, 317–353. Slavica Publishers, Bloomington, Indiana.

[3] Raffaelli, I. (2017) Morfosemantička obilježja percepcijskih glagola u hrvatskom. In: *Zbornik Hrvatskog Slavističkog Kongresa*, 375–387. Hrvatsko filološko društvo.

[4] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

[5] Tadić, M., Fulgosi, S. (2017) Building the Croatian Morphological Lexicon. In: *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, 41–45. Association for Computational Linguistics, Budapest.

The Automatic Translation of Arabic Psychological Verbs Using NooJ Platform

Asmaa Amzali

Computer Science Research Laboratory,
Faculty of Science Ibn Tofail University
Kenitra, Morocco
asmamzali@hotmail.fr

Mohammed Mourchid

Computer Science Research Laboratory
Faculty of Science Ibn Tofail University
Kenitra, Morocco
mourchidm@hotmail.com

Abstract

Arabic is a highly inflected language. Also, it is a morphological and syntactical complex language with differences compared to several more studied languages, like French and English. It may require good pre-processing since it presents important challenges for natural language processing (NLP), especially for machine translation. For this reason, a descriptive and systematic comparative study was carried out with a focus on the inflectional categories of psychological verbs.

The inflectional morphology of these verbs in Arabic is richer and more varied than that of English. Since inflection is the change of word form to mark grammatical distinctions, it occurs in a variety of grammatical classes: nouns, verbs, adjectives, etc.

In this paper, we present a description of the inflectional morphology of psychological verbs in English and Arabic. Since those verbs can be inflected for number, tense, aspect, mood, voice, and agreement. Then we conduct a contrastive analysis of these two languages, which is a systematic study (see Medjdoub, 2022), to identify their differences and similarities. Therefore, such study can be useful in different domains, such as teaching, machine translation, natural language processing, etc.

The objective of our work is to make a contrastive analysis, to clarify the similarities and differences between the two temporal systems of the psychological verbs of both languages, and to specify the corresponding tenses in Arabic. Then, we will create a system of automatic translation of the Arabic psychological verbs using the NooJ platform, where we try to consider the temporal characteristics of each language and solve the problem related to the agreement in gender and number. To realize the automatic translation, we based on our dictionary with about 400 verb entries generated from the lexicon-grammar table of Arabic psychological verbs (Amzali et al., 2019), containing all the lexical, syntactic, semantic, and transformational information of these verbs. Then we will finish by testing the efficiency of this translator on texts and corpora.

Key words

Natural language processing, NooJ platform, contrastive analysis, Arabic psychological verbs, automatic translation

References

[1] Amzali A., Kourtin A., Mourchid M., Mouloudi A., Mbarki S. (2019) Lexicon-Grammar Tables Development for Arabic Psychological Verbs. In: Fehri H., Mesfar S., Silberztein M. (eds.) Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications, 15–26. Springer, Cham.

[2] Cheikhrouhou, H. (2019) Automatic Recognition and Translation of Polysemous Verbs Using the Platform NooJ. In: Fehri, H., Mesfar, S., Silberztein, M. (eds.) Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications, 39–51. Springer, Cham.

[3] Kourtin, A., Amzali, A., Mourchid, M., Mouloudi, A., Mbarki, S. (2019) The Automatic Generation of NooJ Dictionaries from Lexicon-Grammar Tables. In: Fehri H., Mesfar S., Silberztein M. (eds.) Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications, 65–76. Springer, Cham.

[4] Medjdoub, R. (2022) Verb Inflection in English and Arabic: A Contrastive Analysis Study. Milev Journal of Research & Studies, 8(1), 414–422.

[5] Silberztein, M. (2015) La formalisation des langues : l'approche de NooJ. ISTE Editions, London.

Automatic Disambiguation of the Belarusian-Russian Legal Parallel Corpus in NooJ

Valery Varanovich
Mikita Suprunchuk
Yauheniya Zianouka
Yuras Hetsevich
Nastassia Yarash

United Institute of Informatics Problems
Minsk, the Republic of Belarus
ssrlab221@gmail.com

Abstract

Homonymy is still the central problem of automatic text processing at the lexical level. Automatic (rarely semi-automatic) solving of lexical ambiguity was first formulated within the field of science and technology related to the creation of machine translation systems. To date, this is a critical problem of improving the quality of systems for various branches of computational linguistics (Varanovich, 2009).

There are two dominant classes of ambiguity resolution mechanisms (Bouarroudj et al., 2022):

(1) Automatic one, implying a fully computerized solution to this problem.

(2) Interactive (dialogic, semi-automatic) one, supposing a joint solution by a person and a computer.

It means that the user has a set of alternatives from which he should choose one option.

We are conducting work on creating a legal texts corpus in Belarusian and Russian, which is used in various software products (speech synthesis, machine translation, spell checker, etc.). A trilingual Belarusian-Russian-English dictionary of legal terms was created (Hetsevich et al., 2021) within the project. Also, in the process of forming the corpus, contextual dictionaries are compiled, which can become the basis for high-priority dictionaries in the NooJ system. The dictionaries reflect contexts that indicate the preferred translation of a certain term from Russian into Belarusian (Barabash, 2015).

We are planning to create high-priority dictionaries for each of the 26 law codes of Belarus. It is assumed that a lot of diagnostic contexts will be the same in different codes (*настоящий* 'this' = *гэты* 'this' (code), not *настоящий* 'this' = *сапраўдны* 'real' (code)), but we hypothesize that in some cases the contexts for the same values will be different, and it is also possible that in different codes, one polysemous word (or a homonym) will have different meanings.

Here are some words with several meanings which were found during our work: данный – ‘1) гэты, this; 2) дадзены, given’; отпуск – ‘1) выдача (тавараў), issuance, supply; 2) адпачынак (даць), holiday, vacation’. These meanings (translations) could be chosen according to their neighbor words. As NooJ is an effective tool for solving word ambiguity, we plan to use it for compiling syntactic grammars. They will search for homonyms by analyzing the context (the sequence of words) and form a list of various lexical units for different domains. This will assist in identifying terms in different thematic domains, which is very important for compiling special vocabularies for indicated fields.

Key words

Disambiguation, homonym, automatic text processing, dictionary, legal texts corpus

References

[1] Barabash, O. V. (2015) Razgranichieniie omonimii i polisiemii juridichieskikh tierminov (= Барабаш О. В., Разграничение омонимии и полисемии юридических терминов). Rhema. Рема, 2, 39–51. Available at: https://cyberleninka.ru/article/n/razgranichenie-omonimii-i-polisemii-yuridicheskih-terminov.

[2] Bouarroudj, W., Boufaida, Z., Bellatreche, L. (2022) Named Entity Disambiguation in Short Texts over Knowledge Graphs. Knowledge and Information Systems, 64(2), 325–351. Available at: https://doi.org/10.1007/s10115-021-01642-9.

[3] Hetsevich, Y. et al. (2021) Creation of a Legal Domain Corpus for the Belarusian NooJ Module: Texts, Dictionaries, Grammars. In: Bigey, M. et. al. (eds.) 15th International Conference NooJ 2021: Book of Abstracts, 36–37. Available at: http://www.nooj-association.org/nooj2021/Abstracts.pdf.

[4] Silberztein, M. (2003–) NooJ Manual. Available at: http://www.nooj-association.org.

[5] Varanovich, V. V. (2009) Slovar lieksichieskikh valientnostiej v sistiemie russko-bielorusskogo mashinnogo pierievoda (= Воронович, В. В. Словарь лексических валентностей в системе русско-белорусского машинного перевода). In: Третьи чтения, посвященные памяти В. А. Карпова: сб. науч. ст., 108–111. БГУ, Минск. Available at: http://elib.bsu.by/handle/123456789/10645.

Advances in the Automatic Treatment of Newspaper Articles on Economics Journalism Using NooJ

Carmen González

Facultad de Ciencias Económicas y Estadística, UNR,
Rosario, Argentina
caar.gonzalez26@gmail.com

Abstract

Within the field of research of the *Centro de Estudios de Tecnología Educativa y Herramientas Informáticas de Procesamiento del Lenguaje* (CETEHIPL), this work seeks to devise which ways are possible to automatically process a corpus of texts made up by a series of newspaper articles based on economics journalism. The articles were published during the first half of January 2023 in the following newspapers: El Cronista, Ámbito Financiero and El Economista. These newspapers are regarded as newspapers of record in Argentina. An initial pan over the articles reveals present-day issues which are a matter of concern to many Argentine citizens: the external debt, a rising inflation context and the dollar exchange rate. These issues' overriding influence has pushed a number of Argentineans to learn the trade of economists to stay afloat in such a critical setting. In this way, we used NooJ, a linguistic development environment constructed by Max Silberztein, to determine how the press addresses these topics to feed the Spanish-Argentina module (http://www.nooj4nlp.org/resources.html) with specific terminology and enrich our linguistic resources with grammars reflecting the linguistic structures that are typical of the so-called Rioplatense Spanish. We worked with the dictionaries and grammars created by our IES-UNR team to automatically process our corpus.

After studying the corpus, we created in our dictionaries the [+econom] tag to identify specific vocabulary. We found that some terms have been previously included, perhaps, from a general approach while others are totally brand new. Examples include such items as *dólar Qatar* ("Qatar dollar"), *dólar ahorro* ("savings dollar"), *dólar blue* ("blue dollar"), etc. regarding to the scope of dollar exchange rates as well as other finance-related terms such as *inflación ("inflation")*, *deflación ("deflation"), hiperinflación ("hyperinflation"),* or *suba ("increase")* as a noun and not only as a verb. To reflect some specific syntactic structures, we also built grammars for sentences which are very representative of our corpus such as these:

(1) *El 2023 viene seco de dólares. ("2023 is running dry with dollars")*

(2) *Los bonos aguantaron, pero el dólar volvió a calentarse. ("The bonds endured but the US dollar is warming up again")*

(3) *Y en Argentina en particular, el valor del dólar directamente pegó un salto. ("And in Argentina, particularly, the dollar rate leaped").*

These sentences show there is certain degree of personification of the US dollar, which adds extra connotative value to its nominal value due to its disruptive role in Argentina's recent history. Indeed, sociology presents the following perspective:

> *Ask anyone participating in any version of a quiz show in Argentina which was the peso-dollar exchange rate in a not-to-distant date in the past and they can probably tell you that with a fair degree of accuracy. Try, instead, asking who won the 1951 presidential tickets, who the incumbent U.N. Secretary-General is or who the songwriters of a catchy song from the 1960s are and they can probably not tell you so. In other countries, economists or international trade specialists may have these facts and figures ready at their fingertips. In Argentina, however, this information is part of the everyday culture of its of its citizens* (Luzzi and Wilkis, 2019: 14–15; the translation is ours).

Key words

NooJ, automatic treatment, economics journalism, Spanish module Argentina, inflation

References

[1] Ámbito Financiero. Available at: https://www.ambito.com/.
[2] El Cronista. Available at: https://www.cronista.com/.
[3] El Economista. Available at: https://eleconomista.com.ar/.
[4] Luzzi, M. D., Wilkis, A. (2019) El dólar: Historia de una moneda Argentina. *Crítica*, Buenos Aires. Available at: https://planetadelibrospe0.cdnstatics.com/libros_contenido_extra/42/41112_TPCW_El%20dolar. pdf.
[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

Sentiment Analysis of Texts Written in Arabic: Addressing the Issue of Negation

Mohamed El Ammari
Cady Ayyad University
Marrakech, Morocco
elammari1212@gmail.com

Azeddine Rhazi
Cady Ayyad University
Marrakech, Morocco
azeddinrhazi@gmail.com

Salim Rami
Cady Ayyad University
Marrakech, Morocco
Salim.rami@gmail.com

Abstract

Negation is a language-universal feature with a language-specific nature. It is found in all human languages, but the expression of negation varies significantly across languages. Negation was described as a necessity in human communication (Horn, 2001). It has been a topic of inquiry since the early philosophers as Aristotle, and nowadays several research studies have been conducted to account for this complex phenomenon from diverse perspectives: philosophical, logical, cognitive, and linguistic.

As languages differ in the way they express negation, it is found that Moroccan Standard Arabic has a special negation system that is characterized by a high level of metalinguistic complexity in the sense that negation particles in Arabic language carry with them a multitude of referential, metalinguistic, modal, and aspectual meanings (Benmamoun, 2000). Thus, addressing negation in a way that accounts for the peculiarities of the Arabic language system is primordial. The negation system in Arabic is characterized by: (1) association with tense (Kahlaoui, 2019), (2) particle negation is the most common type and (3) six particles constitute the nucleus of the Arabic negation system (lam لم, maa ما, leisa ليس, lammaa لما, laa لا, and lan لن) as in the following examples:

(Sara did not come to work today.)**لم** تحضر سا□ة إلى العمل اليوم.

(You will not reach a solution to your problem when you are angry.) لنْ تصل إلى حل لمشكلتك وأنْتَ غاضب.

(Winter came and it didn’t rain.) أقبل فصل الشتاء ولمّا تمطر السماءُ.

(Ahmed is not lazy.)ليس أحمد متكا□لا.

(Only an ignorant person would say this.) ليس يقو□ هذا إلاّ جاهل.

(No excellence without diligence.)لا تفوق بدو□ اجتهاد.

(Thanks, I don’t smoke.)□كرا، لا أدخن.

(Khalid isn’t present today.) ما خالدٌ حاضرا اليوم.

(You do not do your homework.) أنتم ما تنجزو□ واجباتكم المنزلية.

In sentiment analysis and opinion extraction, negation is a polarity shifter that can change the polarity of an expression, and if not handled well by sentiment analysis

tools, inaccurate results are likely to be generated. This research aims to improve the overall accuracy of sentiment analysis applied on texts written in Modern Standard Arabic language by incorporating a linguistic method of handling negation. To address this issue, this research utilizes linguistic resources provided by the NooJ platform to formalize recognition rules for the Arabic negation system and apply them with a lexicon to a corpus of manually collected data from websites and social media posts. The corpus will be preprocessed before adopting the required NLP analysis. The method starts by describing the rules governing negation in Arabic Language utilizing NooJ lexical and morphosyntactic rules in a lexicon, in addition to local grammar rules in the form of graphs. These steps are specifically designed to recognize negation in texts written in Arabic. The evaluation of the obtained results will serve as a criterion to assess its validity in recognizing negation structures in sentiment analysis in texts written in Modern Standard Arabic. In this respect, it is crucial to carry out an experimentation phase employing the NooJ concordancer.

Key words

Sentiment Analysis, negation, lexicon, NooJ, opinion

References

[1] Benmamoun, E. (2000) *The Feature Structure of Functional Categories: A Comparative Study of Arabic Dialects*. OUP, Oxford & New York.
[2] Horn, L. R. (2001) *A Natural History of Negation*. CSLI Publications, USA.
[3] Kahlaoui, M. H. (2019) Negation in Standard Arabic Revisited: A Corpus-Based Metaoperational Approach. In: Smaïli, K. (ed.) *Arabic Language Processing: From Theory to Practice*, 164–180. Springer, Cham.
[4] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.

A Linguistic Approach for MDU-Based Segmentation

Chahira Lhioui

College of Computing and Information Technology
University of Bisha
Bisha, Saudi Arabia
shahira@ub.edu.sa

Malek Lhioui

Laboratory MIRACL
Sfax, Tunisia
ma.lhioui@gmail.com

Mounir Zrigui

University of Monastir
Monastir, Tunisia
Mounir.Zrigui@fsm.rnu.tn

Abstract

The first step in the analysis of a transcribed speech is its segmentation. It consists of parsing statements into units of a certain type that was previously defined to locate desired information. These units can be at different structural levels involving sentences, clauses or Minimal Discourse Units (MDU), chunks, graphic words, lexical units, morphemes, etc.

Segmentation has been described in several researchs as a crucial stage prior to linguistic treatment. However, discourse segmentation is not so frequent and not taken seriously by most laboratories that treat language automatically. This lack intensifies especially with Arabic language where there is little work on the segmentation of written texts into MDU. In addition, there is virtually no functional and specific clausal segmenter to the Arabic language in the context of an oral spontaneous transcribed conversation because of its ungrammaticalities.

In oral speech, we can have utterances in which two types of MDU can be presented: embedded MDU and overlapping MDU on more than one statement. Indeed, the incomplete and unfinished sentences are principally those which form the dispersed MDU. The following example illustrates scenarios of dispersed MDU on two utterances caused by a non-achieved idea in the first utterance.

Our proposed method is based on the gathering of tokens, obtained during a pretreatment step to form MDU that represent the user's intentions. This is essentially done by using Context-Free Grammar (CFG) rules established from the study of the most relevant oral cases.

To achieve our goals, we opt for NooJ linguistic development platform. This environment provides us with robust and efficient tools and techniques to implement our local grammars.

Key words

NooJ, local grammar, discursive units, context free grammar, segmentation

References

[1] Belguith, H. L., Baccour, L., Mourad G. (2005) Segmentation de textes arabes basée sur l'analyse contextuelle des signess de ponctuations et de certaines

particules. In : Jardino, M. (ed.) *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts (TALN'2005*), 451–456. ATALA, Dourdan.
[2] Iskander, K., Farah, B., Lamia, H. (2013) Segmentation de textes arabes en unités discursives minimales. In : *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 435–449. ATALA, Les Sables d'Olonne.
[3] Keskes I., Benamara F., Belguith L. (2012) Clause-Based Discourse Segmentation of Arabic Texts. In: Calzolari, N. et al. (eds.) *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2826–2832. European Language Resources Association (ELRA).
[4] Mesfar, S. (2008) Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. Thesis. University of Franche-Comté, Besançon.
[5] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.
[6] Silberztein, M., Váradi, T., Tadić, M. (2012) Open-Source Multi-Platform NooJ for NLP. In: Kay, M., Boitet, C. (eds.) *COLING, 24th International Conference on Computational Linguistics, Proceedings of the Conference. Demonstration Papers*, 401-408. Indian Institute of Technology, Bombay.
[7] Touir, A., Mathkour, H., AL-Sanea, W. (2008) Semantic Based Segmentation of Arabic Texts. *Information Technology Journal*, Vol. 7, 1009–1015.

Embracing a Plant-Based Diet: A NooJ Analysis

Isabella Cossidente

University of Salerno
Fisciano, Salerno, Italy
i.cossidente@studenti.unisa.it

Alessandra D'Agostino

University of Salerno
Fisciano, Salerno, Italy
a.dagostino37@studenti.unisa.it

Abstract

This work has been projected and conducted to serve a double purpose. The topic as a whole is the attention towards the vegetarian and vegan diet, a lifestyle which is more and more frequently adopted by people all around the world.

Why do people decide to go vegetarian, even vegan? The first purpose of the work is to answer this question. The Italian population is the chosen sample for the analysis.

The starting point is the Italians' perception of plant-based nutrition, observing the reasons why people decide to follow this kind of diet, and how these same reasons evolved overtime. Specifically, the work covers a ten-year analysis, from 2012 to 2022. The evaluation was made on reports taken from Eurispes, an Italian research institute that publishes annual reports about the main changes in Economics, Politics and Society in Italy.

Three main motives were identified: the most popular one concerns health benefits associated with plant-based nutrition; secondly, there is animal welfare and moral values; the last one is environmental activism. Furthermore, researches have shown that people are pushed towards these diets by the curiosity to try something different.

In working on these first results, NooJ has been used to create syntactic grammars, divided by subject areas, later applied to observe the frequencies and the standard scores of the main concepts. It was possible to reflect on the evolution of the Italian population's approach.

The second aim of the project is to break down stereotypes about the supposed expensiveness and difficulty of plant-based eating.

There are some foods which are notably stigmatized as "vegan foods", i.e. tofu, soy, oat and more. The work wants to show how little these foods can count in a well-balanced diet. An analysis was done examining more than thirty vegan recipes, taken from the blog "Cucina Botanica", a project managed by a well-known Italian content creator, Carlotta Perego, who spreads knowledge on the Internet about veganism and sustainability.

As before, NooJ has been used to construct syntactic grammars, specifically conceived for the refutation of prejudices. The grammars were applied to the mentioned recipes and were used to expose the proportion between "vegan", "unusual" foods on the one hand, and fruits and vegetables, which are more commonly eaten, on the other hand.

Key words

Analysis, diet, NooJ, NooJ syntactic grammars, vegan

References

[1] Cucina Botanica, Ricette, https://www.cucinabotanica.com, last accessed 20/01/2023.
[2] Eurispes: l'Istituto di Ricerca degli Italiani, Rapporto Italia 2012, 2014, 2015, 2017, 2018, 2019, 2020, 2021, 2022, https://eurispes.eu, last accessed 20/01/2023.
[3] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.
[4] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

Differences Between Hate Comments and Insult Comments Directed to Men and Those Directed Towards Women

Martina Galović

Faculty of Humanities and Social Sciences
Zagreb, Croatia
martina.galovic36@gmail.com

Damir Puškarić

Faculty of Humanities and Social Sciences
Zagreb, Croatia
dmrpukari94@gmail.com

Abstract

With the development of the Internet and social networks, it became completely normal to come across harmful comments that insult someone or call for hatred. Even though in the last few years intensive effort has been made on algorithms for hate speech detection, offensive language is still part of everyday life on the Internet. The goal of this paper is to determine if there are differences between hate comments and insult comments directed towards men and those directed at women. To do that, we collected a corpus of around 70 comments for both groups. The collected comments are taken from comment sections on Croatian news portals and Facebook comment sections under the link to the article. To collect comments directed at a certain gender, we looked for the articles about well-known men and women in Croatia, such as singers, actors and actresses, politicians, etc., so we can be sure that each comment is specifically directed toward a man or a woman. For the comment to be included in the corpus, it had to be clear from it that the author of the comment wanted to insult someone, or that he feels hatred towards someone, i.e., we did not take sarcastic comments into account. The corpus was collected in a timespan of a few weeks in December and January of 2022/2023. After collecting it, the spelling errors were corrected, and unknown words were added to the NooJ dictionary. While adding unknown words to the NooJ dictionary, we defined case, gender, and inflectional form for each new word, i.e., we did not add any additional semantic tags to the unknown words. What we are trying to find out with this research is, if there are different syntax patterns used to insult women and to insult men. Also, we want to see if hate speech is more common with one group than another and the same with insult speech.

To be able to do that, first we had to differentiate two types of unwanted comments: hate speech and insulting speech. In the hate speech group, we included comments that aim to propagate intolerance, incite violence or discrimination, and foster prejudice and hatred toward a particular group or person. For the insulting speech group, we chose comments that contain name-calling, mockery, or belittling. The next step was defining several groups of insults and hate speech based on their syntactical structure and creating NooJ syntax grammar diagrams that correspond with each group. The same set of diagrams was used on women and men corpus to see if there are any similarities in comments' structures. In the end, we counted the appearances of each group in collected comments and compared them so we can conclude what kind of syntax structure is more common for women, and what for men. Furthermore, we compared two types of malicious utterances and tried to

understand which of them is more common to appear when talking about women, and which when talking about men. Since hateful and insulting comments are very common in everyday Internet surfing, grammar diagrams like these, and their implementation into machine learning algorithms and connection with usual hateful words (its semantics), can help in recognizing and preventing such comments on social media and news platforms to create more safe and pleasant space.

Key words

Hate comments, insult comments, gender, Internet

References

[1] Kocijan, K., Košković, L., Bajac, P. (2020) Detecting Hate Speech Online: A Case of Croatian. In: Fehri, H., Mesfar, S., Silberztein, M. (eds.) *Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications*, 185–197. Springer, Cham.

[2] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.

[3] Yuan, S., Maronikolakis, A., Schütze, H. (2022) Separating Hate Speech and Offensive Language Classes via Adversarial Debiasing. In: *Proceedings of the Sixth Workshop on Online Abuse and Harms* (WOAH), 1–10.

Explicit Language in English Song Lyrics: Should We Be Worried?

Mila Bikić

Faculty of humanities and social sciences
Zagreb, Croatia
bikic.mila05@gmail.com

Valerija Bočkaj

Faculty of humanities and social sciences
Zagreb, Croatia
vbockaj@gmail.com

Abstract

The usage of explicit lyrics and profane phrases is not a rarity and both male and female artists can be heard using it. Considering factors such as the gender of the artist, the genre, and the year of release, differences in the frequency of explicit language can be observed. In this sense, explicit language includes, but is not limited to verbal insults (i.e. whore, bitch, faggot, the n-word, other racial slurs), linguistic units that are of sexual nature (i.e. dick, coochie, fuck), and other commonly used profane linguistic units (i.e. shitty, bullshit).

To analyze this, syntactical grammar was made in the linguistic development environment software and corpus processor NooJ. First, a corpus consisting of 300 songs was collected. For each of the following six genres – pop, rock, R&B, hip-hop, funk, and country – 50 songs were collected in the following manner: 25 are performed by female artists, and the other 25 by male artists. The songs were chosen from genre charts on the popular music streaming platform Spotify and the lyrics were taken from the website Genius. The chosen songs were the top 50 most popular songs in October 2022 in each genre according to Spotify. The corpus consists of 106 096 tokens – the country genre comprises of 15 270 tokens (8 188 of which were by male artists, 7 082 by female), funk 15 130 (male artists 7 369, female artists 7 761), hip-hop 27 820 (male artists 14 874, female artists 12 946), pop 17 505 (male artists 9 440, female artists 8 065), R&B 16 903 (male artists 8 088, female artists 8 815), rock 13 440 (male artists 6 997, female artists 6 443).

In the second phase, syntactical grammar was designed to recognize and annotate explicit words and phrases in the collected corpus. The grammar recognizes common verbal insults, linguistic units that are of sexual nature, and units that refer to feces. Special attention was given towards phrases that contain the verb fuck and the noun ass, as the notions of rule-bound and rule-breaking linguistic creativity are often applied to these linguistic units. The results were evaluated, giving a precision of 0.95, recall 0.92, and finally, an F-score of 0.93.

As expected, the genre with the highest number of explicit language usage was hip-hop. To our surprise, female hip-hop artists were found to use more profane language than their male counterparts. Finally, a steady incline of explicit words was also observed over following decades: 1950s (consisting of 970 tokens), 1960s (2 371 tokens), 1970s (11 574 tokens), 1980s (5 899 tokens), 1990s (6 192 tokens), 2000s (13 988 tokens), 2010s (29 457 tokens), and 2020s (13 988 tokens). To analyze this, the songs present in the genre corpus were grouped by decade of release. An analysis of this corpus showed that no usage of profane language was present in the 1950s

corpus, however, a total of 512 instances of explicit language was observed in the 2020s corpus.

It can be concluded that the frequency of usage of explicit language tends to grow if the song in question is a hip-hop song and if the song was made in the more recent decades.

Key words

Linguistic analysis, explicit lyrics, music, NooJ, male versus female

References

[1] Chomsky, N. (1957) *Structures*. Mouton, The Hauge.
[2] Mitkov, R., ed. (2004) *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.
[3] O'Keeffe, A., McCarthy, M., eds. (2012) *The Routledge Handbook of Corpus Linguistics*. Routledge, New York.
[4] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.
[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach.* Wiley-ISTE, London.

Comparison of the Representation of Male vs. Female Athletes in Croatian News Portals Using a NooJ Syntax Grammar

Klara Kozolić

Faculty of Humanities and Social Sciences,
University of Zagreb
Zagreb, Croatia
klara.kozolic@gmail.com

Krešimir Štimac

Faculty of Humanities and Social Sciences,
University of Zagreb
Zagreb, Croatia
kkstimac@gmail.com

Abstract

This study investigates the differences in the representation of male and female athletes in Croatian news articles through the analysis of their syntactic constructions. The corpus of 100 articles collected from various news portals featured 57 articles about male athletes and 43 articles about female athletes. The study acknowledges that semantic aspects are crucial to compare the two sub-corpora concerning the question of gender, with the corpus consisting of all the articles found in news portals not directly concerning the objective results and stats of athletes.

The study employed two Croatian syntax grammars in NooJ to examine syntactic structures describing athletes in terms of sport and physical appearance, through "adjective + gender" concordance, where the study observed that syntactic structures containing "<A> + man" showed no concordance, whereas those with "<A> + woman" appeared multiple times. The output of the analysis showed a high level of precision (0.928), moderate recall (0.722), and a good F-score (0.812), also indicating that the syntactic structures used to describe sports and athletes in an objective light were similar or identical for both genders. However, the study found a significant difference in the representation of male and female athletes concerning the emphasis placed on their gender. The analysis revealed that female athletes were more objectified, with 18.48% of all concordance related to their physical appearance, while there was no concordance for male athletes' physical appearance.

This research provides insight into the need for more objective and gender-neutral language in sports journalism. The study's findings suggest that journalists should avoid emphasizing gender and physical appearance when reporting on female athletes. Additionally, sports news portals should be encouraged to use language that reflects the equal status of male and female athletes. In conclusion, this study highlights the importance of understanding the representation of gender in media and the need to eliminate gender bias in sports reporting. Future research could explore other forms of media and sports to provide a more comprehensive understanding of gender representation in sports journalism.

Key words

NooJ, sport, media, Croatian, gender

References

[1] Kocijan, K., Librenjak, S. (2016) Recognizing Verb-Based Croatian Idiomatic MWUs. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.) *Automatic Processing of Natural-Language Electronic Texts with NooJ*, 96–106. Springer, Cham.

[2] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

[3] Silberztein, M., Váradi, T., Tadić, M. (2012) Open-Source Multi-Platform NooJ for NLP. In: Kay, M., Boitet, C. (eds.) *COLING, 24th International Conference on Computational Linguistics, Proceedings of the Conference. Demonstration Papers*, 401-408. Indian Institute of Technology, Bombay.

[4] Šojat, K., Kocijan, K., Bekavac, B. (2018) Identification of Croatian Light Verb Constructions with NooJ. In: Mbarki, S., Mourchid, M., Silberztein, M. (eds.) *Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications*, 96–107. Springer, Cham.

[5] Vučković, K., Tadić, M., Bekavac, B. (2010) Croatian Language Resources for NooJ. *Journal of Computing and Information Technology*, 18(4), 295–301.

Immigrant in the Light of Language Production

Barbara Vodanović

University of Zadar
Zadar, Croatia
bvodanov@unizd.hr

Abstract

We propose to carry out a study of the elements of the co-occurrences of the word "immigrant" in a contrastive Croatian French approach in order to understand the full semantic and axiological potential of language production. The study will be deployed on the lexicographic and journalistic corpus of these two languages.

Our analysis is based on the approach of the Semantics of argumentative possibilities initiated by Galatanu (1999). According to this semantic-discursive theory the language stereotypes represent an open set of associations to the core traits of any word-form. The core traits itself are based upon the lexicological acceptances of word meaning but as Galatanu straits the argumentative possibilities of the core unfold in blocks of external argumentation associating the language production with an element of its stereotype. These associations allow the discursive deployment that constitute the argumentative sequences strictly speaking. We argue that the discursive deployment of core potential is visible in the language production, explicitly in the use of immediate collocates, namely adjectives, of the base immigrant (B) and other contextual co-occurrences organized by linear sequences as: [0,1] prep. + B + [0, 1, 2...]adj. and [0,1] prep.+ adj./n. [0,1,2...]+B for French and for Croatian, as: adj. [0, 1, 2...]+B and B+prep.+NS.

This could unfold the developments of meaning of the word "immigrant" according to the context-based axiology which can be driven both positively and negatively or as non-marked, neutral.

Therefor the paper proposes a syntax grammar in the NooJ tool for recognizing and extracting multi-word units and discontinuous expressions containing the word "immigrant" and gives the results of the application of that grammar on the given (above-mentioned) corpus.

Key words

Immigrant, French, Croatian, argumentative possibilities, discursive deploiement

References

[1] Galatanu, O. (2003) La sémantique des possibles argumentatifs et ses enjeux pour l'analyse de discours. In: Salinero Cascante, M. J., Iñarrea Las Heras, I. (eds.) *El texto como encrucijada: estudios franceses y francófonos,* Vol. 2, 213–226. Universidad de La Rioja, Logroño.

[2] Galatanu, O. (2009) L'Analyse du Discours dans la perspective de la Sémantique des Possibles Argumentatifs : les mécanismes sémantico –

discursifs de construction du sens et de reconstruction de la signification lexicale. In : Garric, N., Longhi, J. (eds.) *L'analyse linguistique de corpus discursifs. Des théories aux pratiques, des pratiques aux théories,* 49–68. Presses universitaires Blaise-Pascal, Clermont-Ferrand.

[3] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.

[4] Silberztein, M. (2015) La formalisation des langues : l'approche de NooJ. ISTE Editions, London.

[5] Virgine, M. (2009) La Sémantique des Possibles Argumentatifs : un modèle de description-construction-représentation des significations lexicales. *Cahiers de Narratologie* [en ligne], 17. Available at: http://journals.openedition.org/narratologie/1337.

Semantic Analysis of Migrants' Self-Entrepreneurship Ecosystem Narratives

Cecilia Olivieri

X23 Science in Society
Treviglio, Italy
cecilia.olivieri@x-23.org

Lorenzo Maggio Laquidara

X23 Science in Society
Treviglio, Italy
lorenzo.maggio@x-23.org

Jie Sheng

X23 Science in Society
Treviglio, Italy
jie.sheng@x-23.org

Agathe Semlali

X23 Science in Society
Treviglio, Italy
agathe.semlali@x-23.org

Abstract

Migrant entrepreneurship in Europe constitutes a multifaceted landscape, characterized by the interactions of a rich array of actors with diverse ambitions and needs. In spite of its sociological, political, and civic relevance, this topic has not received enough academic attention, and approximative narratives on migrant entrepreneurship are widespread among institutional actors. Mainstream media narratives often portray migrants as a vulnerable social group claiming financial, material, and labor resources from arrival countries' public. At best, media actors justify the need for socio-economic integration based on the immigrants' special needs and often dire living conditions – some narrative referred to as "miserabilistic" (Desille and Nikielska-Sekula, 2021). Other than running the risk of fostering the sense of alterity between locals and migrants, normally these narratives "homologate" public perception on migrants. Migrant flows articulate themselves in multiple generations, places of origin, demographic compositions, and purposes. All these "axes of difference" interact to form complex ecologies, especially in urban environments. Therefore, self-perception directly collected from the field represents a valuable tool to navigate migrant communities in Europe: such is true for migrants and locals alike, given a common interest in defining themselves and the spaces they occupy (Buhr, 2021).

Migrant entrepreneurs represent crucial mediators between exogenous and self-produced narratives. Because of their ability to generate economic and social wealth, migrant entrepreneurs openly challenge social stigmatization. They also elicit cultural enrichment by contributing their migration background to the local entrepreneurial panorama. More importantly, their privileged position offers them a vantage point to describe their personal experiences with migration (Zanfrini, 2015), their communities, their approach to the entrepreneurship ecosystem (the market itself, as well as the actors on the ground, e.g. incubators, mentors and coaches, experts and scientists, economic traders, etc.), within the countries of destination (Buhr, 2021).

In this context, our study wishes to investigate the migrant entrepreneurship environment and its stakeholders (business incubators, research experts, policymakers, migrant entrepreneurs themselves, and a variety of other figures), advancing a transversal, multi-source, and multi-level narrative. Migrant entrepreneurs view their pursuits as significant achievements for personal and communal wellbeing. Policymakers, support-program designers, *third-sector* practitioners, and private investors, on the contrary, often perceive (and reportedly say about) migrant entrepreneurship as inapt for competition in the local entrepreneurial environment. These perceptions and narratives perform at different levels the actions of individuals, through the constant refinement of rigid, impeding, and very difficult to deconstruct phenotypic codes. One of the objectives of our work is precisely that of, starting from language, investigating these codes, defining their weight, connections, and reciprocal influences, formalizing them, and studying them for their overcoming.

We identified individual, semi-structured interviews as the chief instrument to explore these competing narrative perspectives. We intend to process these transcribed natural-language inputs using NooJ, especially co-occurrence sentiment analysis to reveal the role of community in migrant self-entrepreneurship in the EU. Corpus-level discourse analysis will be first supported by lexical and semantic concept formalization in NooJ vocabularies and then by NooJ grammars, able to perform context disambiguation. Syntax grammars will serve as our primary tool in sentiment analysis. We especially consider prepositions and location adverbs as markers for the concept of *community*. In fact, our study rests on the assumption that preposition-based syntax grammars can reveal attitudes and dispositions within a text (Monteleone, 2019). Our interviews are far from all being "success stories", and most of them dilute the respondents' general views on being entrepreneur in Europe with personal stories. In summary, our study intends to unveil a holistic narrative on migrant entrepreneurship in Europe that is capable of accurately describing its complex articulation in at the individual, community, professional, and policy levels.

Key words

Migrant self-entrepreneurship, sentiment analysis, ecosystem, phenotypic codes, public narrative

References

[1] Buhr, F. (2021) Migrants Mental Maps: Unpacking Inhabitants Practical Knowledges in Lisbon. In: Nikielska-Sekula, K., Desille, A. (eds) *Visual Methodology in Migration Studies*, 51–67. Springer, Cham.

[2] Desille, A., Nikielska-Sekula, K., eds. (2021) *Visual Methodology in Migration Studies*. 1st edn. Springer, Cham.

[3] Monteleone, M. (2019) *NooJ Grammars and Ethical Algorithms*. Università degli Studi di Salerno, Fisciano.

[4] Silberztein, M., Monti, J., Monteleone, M., eds. (2014) *Formalizing Natural Languages with NooJ*. Cambridge Scholars Publishing, Newcastle upon Tyne.

[5] Zanfrini, L. (2015) *The Diversity Value: How to Reinvent the European Approach to Immigration*. McGraw-Hill Education, Maidenhead.

Reviewing the Position of the “Other” in Croatia’s “Non-European” Collections in the Second Half of the 20th Century Using NLP

Martina Bobinac

Institute of Art History
Zagreb, Croatia
mbobinac@ipu.hr

Abstract

Ethnographic museums lie on the legacy of colonialism. The first ethnographic collections were based on former cabinets of curiosity where the aim was to collect “exotic”, “different” and “primitive” cultures of “otherness”, usually coming from West’s colonies, and to present them to the Western public. They were presented in colonial museums and world fairs, where entire villages of people were brought and caged, exhibited as “exotic artefacts”. The second half of the twentieth century brought a change in this perspective- the postcolonial theory was born, and with it, the notion of the „other” was coined. It is a critical narrative that indicates the tendency to describe someone else’s culture, society, object, or social group as “different”, “distant”, “strange” or “external” from the perspective of the society from which the speaker is coming from. One of the biggest challenges of European colonial heritage was reinventing and adjusting ethnographic, but especially colonial museums to contemporary theoretical approaches, since it was necessary to present non-European cultures in a postcolonial light.

This work will try to follow and articulate this change in perspective on the example of the Ethnographic Museum of Zagreb. In 1919, the Ethnographic Museum in Zagreb was founded. Its World Cultures collection, historically called The Collection of Non-European Cultures, holds over 3 000 artifacts and works of art. The items featured in the collection originate from South America, Africa, Asia and Australia. Using archival data that bring professional, journalistic, and public perceptions of this collection from 1950 to 1990, the purpose of the paper is to compare the perception of the “other” in Croatian public and cultural sphere over time. The dataset was created by applying OCR using Python to all available newspaper articles and articles from other specialized journals in the timespan of 40 years (from 1950 to 1990) on the topic of “non-European” artifacts and art in Croatian collections. The aforementioned archival data was collected in Ethnographic museums of Zagreb and Split, Croatia. By applying Natural Language Processing on this dataset in Python and Spacy, the difference in discourse and rhetoric is examined with results that confirm that a shift in perception occurred over time, where the development of global consciousness on topics of decolonization is reflected in Croatian collections and related publications. This change in discourse and rhetoric was created using sentiment analysis by creating a list of what is historically recognized as colonial discourse, labeling those words (nouns, adjectives, and word clusters) as “negative” and a list of what is recognized as discourse connected to decolonial theory, labeling those words as “positive”. By implementing these lists of “negative” and “positive”

words on the whole dataset, a sentiment analysis of texts through time is examined. The results are presented by using graphs in Python which clearly show a rise in the number of "positive" and a decline of "negative" words as time progresses.

Key words

Ethnographic museums, colonialism, Croatia, decoloniality, colonial discourse

References

[1] Chen, Y., Khoury, A. (2021) Decolonisation of Past and Present Identities: A Discussion on the Representations of 'Britishness' and 'Otherness' in UK Museums. In: *Proceedings of the 7th International Conference on Humanities and Social Sciences Research,* 954–957. Atlantic Press.

[2] Mignolo, W. D., Walsh, C. E. (2018) *On Decoloniality: Concepts, Analytics, Praxis*. Duke University Press.

[3] Nebbou, A. (2013) The Colonial Discourse Versus the Anti-Colonial. *Scholars World* 1(3), 24–29.

[4] Newton, K. M., Homi, K. B. (1997) The Other Question: The Stereotype and Colonial Discourse. In: Newton, K. M. (ed.) *Twentieth-Century Literary Theory*, 293–330. Palgrave, London.

[5] Sladojević, Ana. (2011) Muzej kao slika sveta: prostor reprezentacije identiteta i ideologije. Doctoral dissertation. Univerzitet umetnosti u Beogradu, Interdisciplinarne studije, Beograd.

Engaging with the Agenda 2030

Stella Nunzia Costanza
University of Salerno
Fisciano, Salerno, Italy
s.costanza1@studenti.unisa.it

Antonio Pagano
University of Salerno
Fisciano, Salerno, Italy
a.pagano86@studenti.unisa.it

Antonio Duca
University of Salerno
Fisciano, Salerno, Italy
aduca561@gmail.com

Abstract

This project aims to analyze, through the use of NooJ, the study of the United Nations' 2030 Agenda for Sustainable Development program. The action plan for people, planet, and prosperity was signed on September 25th, 2015 by 193 United Nations member countries, including Italy, to share the commitment to ensure a better present and future for our planet and the people who inhabit it. The Global Agenda defines 17 Sustainable Development Goals to be achieved by 2030, articulated in 169 targets, which serve as a compass to put Italy and the world on a sustainable path. The process of changing the development model is monitored through the goals, targets, and over 240 indicators: based on these parameters, each country is periodically evaluated by the United Nations and national and international public opinion. The 2030 Agenda brings with it a great novelty: for the first time, a clear judgment is expressed on the unsustainability of the current development model, not only on the environmental level but also on the economic and social levels, thus definitively overcoming the idea that sustainability is solely an environmental issue and affirming an integrated vision of the various dimensions of development.

However, there are differences between territories in Italy: we currently live in a two-speed country, with significant disparities, for example, on commitments to meet the goals of the UN Agenda 2030. The new "Territories Report" of the Italian Alliance for Sustainable Development (ASviS) tells how this commitment changes according to the regions and geographical areas of Italy. Through statistical indicators and regional data, the report attempts to analyze the positioning of provinces, metropolitan cities, and urban areas in relation to the now well-known 17 Sustainable Development Goals of the United Nations. What emerges is a portrait of an Italy "at different speeds, where territorial differences increase rather than decrease."

The analysis of texts and articles through the NooJ software, thanks to grammar constructions, statistics, and frequencies, shows which possible proposals or solutions are highlighted to improve or offer solutions to extinguish the problem. Through the analysis of various texts, the ideologies from different perspectives that are being implemented and will be implemented for the good of our planet can be observed.

Key words

NooJ, Agenda 2030, ecosystem, planet, grammars

References

[1] Gli eventi estremi in l'Italia a causa del clima, https://tg24.sky.it/ambiente/2022/12/30/clima-eventi-estremi-italia-2022?social=facebook_skytg24_photo_null, last accessed 30/01/2023.

[2] Gli obiettivi dell'Agenda 2030, https://www.lasvolta.it/4807/italia-ancora-lontana-dagli-obiettivi-dellagenda-2030, last accessed 19/01/2023.

[3] Lotta al cambiamento climatico, https://asvis.it/l-agenda-2030-dell-onu-per-lo-sviluppo-sostenibile/#, last accessed 2023/01/21.

[4] Silberztein, M. (2003–) *NooJ Manual*. Available at: www.nooj4nlp.org.

[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

NATURAL LANGUAGE PROCESSING APPLICATIONS

NooJ Grammars for Morphophonemic Continuity and Semantic Discontinuity Title

Mario Monteleone

Università degli Studi di Salerno
Fisciano, Salerno, Italy
mmonteleone@unisa.it

Abstract

Morphophonemic Continuity (MC) makes it possible to connect the meaning of two or more words if they all include the same lexical morpheme (LM). Often, but not always, these words may belong to different grammatical categories, based on the phonemes/allophones/morphemes/allomorphs that co-occur inside them with the lexical morpheme in common. For instance, the Italian lexical morpheme *infett-* connects the two verbs *infettare* (to infect) and *disinfettare* (to disinfect), but also other words as *infettivologo* (infectious disease specialist, noun), *infettivo* (infectious, adjective) and *disinfettante* (disinfectant, noun, and adjective).

As known, MC is at the basis of the notion of support verb/support verb extension and predicative nouns/adjectives established by Maurice Gross' Lexicon-Grammar (LG). In any language, support verbs do not have a predicative function, and in any sentences, they neither select arguments nor participate in meaning definition. Actually, their only role is to "support" the function of nominal/adjectival predicates, which are connected by MC to ordinary verbs, as for instance in Max adora Maria = Max ha adorazione per Maria (Max adores Maria = Max has adoration for Maria). In such cases, MC connects the meanings of Ordinary Verb Sentences (OVSs), which are synthetic, to those of Support Verb Constructions (SVCs), which are analytic.

However, in Italian, even when a LM is in co-presence and preserved, some SVCs have a meaning quite significantly different from that of their corresponding OVSs, that is to say: they succeed in preserving morphophonemic continuity but fail in preserving a complete semantic continuity. This paper wants to analyze this phenomenon, which we will see is rather regular if we take into consideration a specific class of nouns and specific uses of determiners/predeterminers. The nouns that create this "semantic anomaly" express the action of "touching/hitting someone with a body part or an object", as for instance *bastonata* (blow struck with a stick), *manata* (blow struck with a hand), *spintone* (shove), *racchettata* (blow struck with a racket), and so on. The co-occurrence of these nouns regularly creates sentences that have only the structure of SVCs, but which are very often OVSs, and in some case Idiomatic Sentences (ISs) to lexicalize as Atomic Linguistic Units (ALUs). Specifically, we will see how inside certain seeming SVCs, noun grammatical number is at the base of this "semantic anomaly", manly when the verbs occurring in the corresponding OVSs endogenously provide for a plural number, i.e. from two onwards. Therefore, we will build specific NooJ grammars to morphosnitactically describe and address this "semantic anomaly", providing ad hoc examples and

documenting, where necessary, the different syntactic patterns by which the correspondence between OVSs and SVCs is satisfied or fails.

Key words

NooJ, NooJ grammars, morphophonemic continuity, semantic discontinuity

References

[1] Gross, M. (1975) *Méthodes en syntaxe*. Hermann, Paris.
[2] Gross, M. (1981) Les bases empiriques de la notion de prédicat sémantique. *Langages*, 15(63), 7–52.
[3] Gross, M. (1990) Sur la notion harrissienne de transformation et son application au français. *Langages*, 25(99), 39–56.
[4] Silberztein, M. (2003–) *NooJ Manual*. Available at: http://www.nooj-association.org.
[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

Exploring Digital Literary Communication: Theoretical, Processual, and Environmental Perspectives with a Case Study of Linguistic Analysis and Graphical Representation of Dante Alighieri's Italian Language

Francesco Saverio Tortoriello

University of Salerno, Italy
fstortoriello@unisa.it

Ilaria Veronesi

University of Salerno, Italy
iveronesei@unisa.it

Ritamaria Bucciarelli

University of Siena, Italy
ritamaria.bucciarelli@unisi.it

Andrea Rodrigo

University of Rosario, Argentina
andreafrodrigo@yahoo.com.ar

Abstract

As technology evolves, a wide variety of languages have enabled us to create new methods of communication which implement formalized codes for intelligent communication, such as syntactically homologated systems. The proliferation of acronyms in the digital realm demonstrates the ability of humans to translate, reformulate, and convert images, sounds, and feelings into a multicodical formal language. Algorithms and acronyms have become a substitute for words and phrases, with the virtual communicator 'Mood' being the symbol of our society. We are inevitably steered towards digital codes, which carry predefined semantics, generated by a fixed sentence.

The research encompasses multiple projects and focuses on literary communication in digital, an operational hypothesis that has been validated by multiple scientific reference models to confirm its validity. The primary goal of the entire research is the process of elaboration of Natural Language Processing (NLP) capabilities. The proposed model takes into account the process of understanding, analyzing and validating the data retrieval, in order to realize a formal process that goes from textual prediction to the first formal process of production of linguistic validation data.

This work seeks to outline an irrefutable logical process, encompassing normative grammar and formal grammar, to apply the poetic traits observed in Dante Alighieri's Divine Comedy across different stages:

(1) We will analyze Dante Alighieri's *Divine Comedy* with sentiment analysis to identify rhetorical figures and metrics in the sections being studied.

(2) We will then transfer the linguistic environment to NooJ to evaluate formal mechanisms and validations with distributive and transformational analysis, which will be displayed in graphs and Peco tables. We will also compare and detect traits of the Divine Comedy with a phonetic analysis and Silvestri's rhetoric to study recursive processes.

(3) We will then explore implementation in Digital intelligence Word Tool.

(4) Finally, we will analyze scientific validation, quantum theories, and verses to highlight emotional traits in the Fano Plan.

M. Planat's quantum hypothesis proposes a scientific process of transforming poetic language into a mathematical sequence, as well as a formal code. Linguists specializing in lexico-grammar grammars provide an NLP transcription of the author's characteristics into formal language, describe in Lexicon-Grammar, and then create the necessary linguistic resources in the NOOJ environment, consisting of graphs to validate the Dante tercet. With integration of the corpus and local grammars, this can be then reformulated and translated.

Within the scope of our research, Ahmida Bendjoudit's (2020) model has the capacity to detect the novel parameters of linguistic production, such as tokens, etc. The characteristics picked are the consequence of indisputable procedures, the emotional features of fixed structures in experimental linguistic equations of both first and second level. Max Silberztein's (2015) NooJ system advances the production of analysis and paraphrasing of sentences, as well as tools to construct formal dictionaries and grammatical and NLP applications, such as automated semantic annotators and paraphrase generators.

Digital intelligence W.T. enables the generation of fixed sentence analysis and output in high-performance computing settings to generate textual paraphrases. It provides management, processing, and retrieval capabilities along with advanced statistical analysis through data collection techniques applied to images of created components and metrics for image measurement.

Key words

Algorithms, acronyms, digital codes, poetic text, Fano Plane

References

[1] Gross, M. (1975) *Méthodes en syntaxe*. Hermann, Paris.
[2] Planat, M., Saniga, M., Kibler, M. R. (2006) Quantum Entanglement and Projective Ring Geometry. *SIGMA*, Vol. 2, 066, 1–14.
[3] Planat, M., Saniga, M. (2007) On the Pauli Graphs of N-Qudits. arXiv preprint quant-ph/0701211.
[4] Silberztein, M. (1999) Text indexation with INTEX. *Computers and the Humanities*, 33(3), 265–280.
[5] Silberztein, M. (2016) *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE, London.

Machine Translation and Multi-Words Expressions: Pragmatic Analyzers Techniques Using NOOJ

Ali Boulaalam

Moulay Ismail University
Meknes, Morocco
lingdroit@gmail.com

Nisrine El Hannach

Mohamed First University
Oujda, Morocco
nisrine.elhannach81@gmail.com

Abstract

This contribution is a presentation of a project being developed by a team of linguists addressing machine translation of multi-word expressions in relation to natural languages processing NLP, as being a linguistic phenomenon of great importance in the field of computational linguistics. The presentation will highlight the development of this machine translation system based on methodologies and technics of linguistic engineering that inevitably enable the construction of a pragmatic analyzer that regulates its semantic aspects. This requires leveraging elements of platform linguistics, which rely on theoretical frameworks, formal methodologies, and effective computer environments. To achieve this, a hybrid theoretical framework that combines contemporary linguistic theories with computational formalism in the development of an operable machine translation system. NooJ environment, an open-source computer platform, was adopted to conduct applied translation operations of multi-word expressions from Arabic to French and English and the project is planning to include more languages.

Key words

Machine translation, platform linguistics, pragmatic analyzer, NOOJ platform, multi-word expressions

References

[1] Salah, M. (2008) Figement et traduction : problématique générale. *Meta,* 53(2), *244-252*

[2] Silberztein, M. (2015) *La formalisation des langues : l'approche de NooJ.* ISTE Editions, London.

[3] Silberztein, M. (2019) Les outils informatiques au service des linguistes présentation. *Langue française*, 203(3), 7–14.

[4] • محمد الحناش: المعجم الآلي اللغة العربية: قاعدة بيانات للتعبيرات المسكوكة، مجلة التواصل اللساني، عدد خاص 2009.

[5] • محمود اسماعيل الصيني: الترجمة الآلية واللغة العربية، مجلة التواصل اللساني، المجلد السادس، العددان 1و2.

[6] مجدي حاج ابراهيم – عائشة رابح محمد: نظم الترجمة الإحصائية والتحويلية، دراسة تحليلية مقارنة، مجلة الدراسات اللغوية والأدبية المجلد الثالث العدد الأول

[7] • عبدالله بن حمد الحميدان: مقدمة في الترجمة الآلية مكتبة العبيكان 2001

[8] علي بولعلام، لسانيات المنصات واللغة العربية، تقنيات استخدام بيئة نوج، (2018)


International NooJ2023 Conference
May 31 - June 2, Zadar, Croatia