

14TH International conference NOOJ 2020 : book of abstracts

Edited book / Urednička knjiga

Publication status / Verzija rada: **Published version / Objavljena verzija rada (izdavačev PDF)**

Publication year / Godina izdavanja: **2020**

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:131:006205>

<https://doi.org/10.17234/9789531758727>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-07-10**



Sveučilište u Zagrebu
Filozofski fakultet
University of Zagreb
Faculty of Humanities
and Social Sciences

Repository / Repozitorij:

[ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences](#)





14TH INTERNATIONAL CONFERENCE NOOJ 2020

Book of Abstracts

Virtual Conference
Zagreb, Croatia
June 5-7, 2020

Božo Bekavac
Kristina Kocijan
Max Silberztein
Krešimir Šojat (Eds.)

Editors

Božo Bekavac

*Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, 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*

Krešimir Šojat

*Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia*

Proofreading

Petra Bajac

ISBN

ISBN 978-953-175-872-7

DOI



10.17234/9789531758727

Publisher

Filozofski fakultet Sveučilišta u Zagrebu - FF Press

FF press

Zagreb, 2020.

Organization

14th Annual NooJ 2020 International Conference was organized as a Virtual conference.

Organizing Institutions



*Department of Information and Communication Sciences
and Department of Linguistics*
Faculty of Humanities and Social Sciences in Zagreb, Croatia



ELLIADD Laboratory



NooJ Association

Organizing Committee

Kristina Kocijan	University of Zagreb, Croatia
Krešimir Šojat	University of Zagreb, Croatia
Božo Bekavac	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
Farida Aoughlis Mouloud Mammeri University, Algeria
Božo Bekavac University of Zagreb, Croatia
Vincent Bénét Centre d'Etudes Franco Russe CNRS, Russia
Xavier Blanco Autonomous University of Barcelona, Spain
Mohamed El Hannache Sidi Mohamed Ben Abdellah University, Fes, Morocco
Héla Fehri University of Sfax, Tunisia
Zoe Gavriilidou Democritus Univ. of Thrace, Greece
Yuras Hetsevich National Academy of Sciences, Belarus
Kristina Kocijan University of Zagreb, Croatia
Philippe Lambert Université de Lorraine, France
Denis Le Pesant Université Paris 10, France
Peter Machonis Florida International University, USA

Samir Mbarki IbnTofail University, Morocco
Slim Mesfar University of Manouba, Tunisia
Elisabeth Métais Conservatoire National des Arts et Métiers, France
Mario Monteleone University of Salerno, Italy
Johanna Monti University of Naples, Italy
Kamal Naït-Zerrad INALCO, France
Thierry Poibeau Laboratoire Lattice, CNRS, France
Jan Radimský University of South Bohemia, Czech Republic
Andrea Rodrigo University of Rosario, Argentina
Krešimir Šojat University of Zagreb, Croatia
François Trouilleux Université Blaise-Pascal, France

Preface

Preface

NooJ is a linguistic development environment that provides tools for linguists to construct linguistic resources that formalize a large gamut of linguistic phenomena: typography, orthography, lexicons for simple words, multiword units and discontinuous expressions, inflectional, derivational and agglutinative morphology, local, phrase-structure and dependency grammars, as well as transformational and semantic grammars. For each linguistic phenomenon to be described, NooJ proposes a set of computational formalisms, the power of which ranges from very efficient finite-state automata (that process regular grammars) to very powerful Turing machines (that process unrestricted grammars). NooJ also contains a rich toolbox that allows linguists to construct, maintain, test, debug, accumulate and share linguistic resources. This makes NooJ's approach different from most other computational linguistic tools that typically offer a unique formalism to their users and are not compatible with each other.

NooJ provides parsers that can apply any set of linguistic resources to any corpus of texts, to extract examples or counter examples, annotate matching sequences, perform statistical analyses, and so on. Because NooJ's linguistic resources are neutral, they can also be used by NooJ's generators to produce texts. By combining NooJ's parsers and generators, one can construct sophisticated NLP (Natural Language Processing) applications such as MT (Machine Translation) systems, abstracts and paraphrases generators, etc.

Since its first release in 2002, several private companies have used NooJ's linguistic engine to construct business applications in several domains, from Business Intelligence to Opinion Analysis. To date, there are NooJ modules available for over 30 languages; more than 140,000 copies of NooJ have been downloaded.

NooJ has also been enhanced with new features to respond to the needs of researchers who need to analyze texts in various domains of Human and Social Sciences (history, literature and political studies, psychology, sociology, etc.). Since 2013, a new version for NooJ is available, based on the JAVA technology and available to all as an open source GPL project and distributed by the European Metashare platform.

This volume contains abstracts of the 42 papers presented at the International NooJ 2020 conference, which was organized as a virtual conference by the Faculty of Humanities and Social Sciences in Zagreb, the University of Franche-Comté and the NooJ International Association. The Book of Abstracts starts with the abstracts of two invited speakers,

- Simon Krek: *Digital Dictionary Databases and ELEXIS Dictionary Matrix*;
- Anita Peti-Stantić: *Data to Help us Form Images: the Abstract and Non-Imageable*.

The other presentations are organized in seven sessions.

The first session is dedicated to the Digital Humanities:

- Max Silberztein: *NooJ for the Digital Humanities*;
- Peter Machonis: *NooJ in the Humanities: Phrasal Verb Usage in the Works of British and American Authors*;
- Piton Odile: *Maria Deraismes' Writings on Women. Study Using NooJ Linguistic Platform*;
- Lorena Kasunić and Gordana Kiseljak: *Depictions of Women in Croatian "Duga" and "Tena" - a Computational Analysis*;
- Raffaele Manna, Antonio Pascucci, Maria Pia Di Buono and Johanna Monti: *The Use of Figurative Language in a Dream Descriptions Italian Corpus: Exploiting NooJ for Stylometric Purposes*;
- Cristina Mota, Diana Santos and Anabela Barreiro: *Paraphrasing Emotions in Portuguese*;
- Lesia Kaigorodova: *Text Analysis of Scientific Papers Using NooJ: Case Study with Vitamin D Supplementation Debates*;
- Tong Yang: *Automatic Extraction of French Food Expression Routine with NooJ*.

The second session is dedicated to Multiword Expressions and Named Entities:

- Kristina Kocijan, Krešimir Šojat and Silvia Kurolt: *Multiword Expressions in the Croatian Medical Domain: Who Carries the Domain Specific Meaning?*
- Aleksandar Petrovski: *Named Entity Recognition with NooJ*;
- Diego Válio Antunes Alves, Božo Bekavac and Marko Tadić: *Optimization of Portuguese Named Entity Recognition and*

Classification by Combining Local Grammars and Conditional Random Fields Trained with Parsed Corpora;

- Roua Torjmen and Kais Haddar: *Automatic Recognition and Translation of Tunisian Dialect Named Entities into Modern Standard Arabic.*

The third session is dedicated to Text Mining and Question Answering:

- Essia Bessaies, Slim Mesfar and Henda Ben Ghazela: *Annotation of Cause-Result Questions in Standard Arabic Using Syntactic Grammars*
- Kaoutar Belhoucine, Mohammed Mouchid, Aziz Mouloudi and Samir Mbarki: *A Bottom-up Approach for Moroccan Legal Ontology Learning from Arabic Texts;*
- Sondes Dardour, H la Fehri and Kais Haddar: *Development of a Question-Answering System in the Legal Field Based on Ontologies;*
- Ismahane Kourtin and Aziz Mouloudi and Samir Mbarki: *Development of a Question-Answering System in the Legal Field Based on Ontologies.*

The fourth session is dedicated to AI, Lexicons and Dictionaries:

- Mario Monteleone: *NooJ for Artificial Intelligence: an Anthropic Approach;*
- Raffaele Marcone, Rosa Giulia, Roberto Capone, Giulia Savarese, Marianna Greco, Colomba La Ragione and Janvier Julian Enriquez: *New Global Cloud Solutions and NooJ's Woven Digital Intelligences for Homologated Synthetic Fixed Communication;*
- Asmaa Kourtin, Mohammed Mouchid, Abdelaziz Mouloudi and Samir Mbarki: *Standardization and Implementation of Lexicon-Grammar Tables in NooJ Platform.*
- Katarina Aladrovi  Slova ek: *Using Linguistic Software NooJ in Describing Preschool and Younger School Children's Croatian Language Vocabulary;*
- Vincent B net: *New Russian Resources for Silberztein's Software NooJ;*
- Mohamed El Hannach and Ahmed Bounoua: *Al-Erfan-DIC: The Electronic Dictionary of Standard Arabic Using NooJ Platform;*
- Annibale Elia, Alessandro Maisto, Lorenza Melillo and Serena Pelosi: *Lexical Complexity and Basic Vocabulary of the Italian Language;*
- Linda Miji  and Anita Bartulovi : *Formalizing the Latin Language on the Example of Medieval Latin Wills;*
- Mourad Aouini and Laure-Anne Caraty: *Morphology of Middle French Verbs with NooJ;*

- Lena Papadopoulou and Elina Chatjipapa: *A Morphological Grammar for Modern Greek: State of the Art, Evaluation and Upgrade*.

The fifth session is dedicated to to Syntax and Semantics:

- Gaurish Thakkar, Nives Mikelić Preradović and Jeremy Barnes: *Detecting Negation Scope with NooJ*;
- Prihantoro: *The Annotations of Indonesian Reduplications with NooJ*;
- Ali Boulaalam and Azeddine Rhazi: *Arabic Transformational Based Approach: the Automatic Paraphrasing of Syntactic Structures*;
- Magaly Bigey: *Using NooJ for Marketing Choices*;
- Walter Koza and Hazel Barahona: *Computational Modeling of a Nominal Ellipsis Grammar for Spanish*.

The sixth session is dedicated to Teaching with NooJ:

- Andrea Rodrigo, Silvia Reyes and María Andrea Fernández Gallino: *Automatic Treatment of Causal, Consecutive and Counterargumentative Discourse Connectors in Spanish: a Pedagogical Application of NooJ*;
- Mirela Landsman Vinković and Kristina Kocijan: *Preparing the NooJ German Module for the Analysis of a Learner Spoken Corpus*.

The seventh session is dedicated to the description of Verbs:

- Olena Saint-Joanis: *Formalizing Ukrainian Verbs with NooJ*;
- Maximiliano Duran: *Transformations and Paraphrases of Quechua Sentiment Predicates*;
- Masako Watabe: *Verbal syntax in Rromani: Diatheses*;
- Asmaa Amzali, Mohammed Mourchid, Abdelaaziz Mouloudi and Samir Mbarki: *Arabic Psychological Verbs Recognition through NooJ Transformational Grammars*;
- Hamid Annouz: *Formal Processing of Values of the Simple Kabyle Aorist Using NooJ*;
- Ilham Blanchete, Mohammed Mourchid, Samir Mbarki and Abdelaziz Mouloudi: *Formalizing Arabic Deverbal Noun "Gerund" Using NooJ Platform*.

This volume should be of interest to all users of the NooJ software because it presents the latest development of its linguistic resources as well as a large variety of applications. Linguists as well as Computational Linguists who work on Arabic, Belarus, Croatian, English (*American English, Late Modern English, Old English*), French, German, Greek, Indonesian, Italian, Kabyle, Latin, Macedonian, Portuguese, Rromani,

Russian, Quechua, Spanish or Ukrainian will find advanced up-to-the-minute linguistic studies for these languages.

We think that the reader will appreciate the importance of this volume, both for the intrinsic value of each linguistic formalization and the underlying methodology, as well as for the potential for developing NLP applications along with linguistic-based corpus processors in the Social Sciences.

The Editors

Contents

Editors.....	1
Publisher.....	1
ORGANIZATION	i
Organizing Institutions	i
Organizing Committee.....	i
Scientific Committee	i
PREFACE	iii
Editors' Preface.....	iv
Contents	ix
INVITED TALKS.....	1
Digital Dictionary Database and ELEXIS Dictionary Matrix	
Simon Krek	2
Data to Help us Form Images: the Abstract and Non-Imageable	
Anita Peti-Stantić.....	4
1. DIGITAL HUMANITIES	7
NooJ for the Digital Humanities	
Max Silberztein	8
NooJ in the Humanities: Phrasal Verb Usage in the Works of British and American Authors	
Peter Machonis.....	10
Maria Deraismes' Writings on Women: Study Using NooJ Linguistic Platform	
Oditel Piton	12
Depictions of Women in “Duga” and “Tena” - a Computational Analysis	
Lorena Kasunić and Gordana Kiseljak.....	14
The Use of Figurative Language in a Dream Descriptions Corpus. Exploiting NooJ for Stylometric Purposes	
Raffaele Manna, Antonio Pascucci, Maria Pia di Buono and Johanna Monti	16

Paraphrasing Emotions in Portuguese	
Cristina Mota, Diana Santos and Anabela Barreiro	18
Text Analysis of Scientific Papers using NooJ: Case Study with Vitamin D Supplementation Debate	
Lesia Kaigorodova	20
Automatic Extraction of French Foods Expression Routine with NooJ	
Tong Yang	22
2. MULTI WORD EXPRESSIONS AND NAMED ENTITES.....	25
Multiword Expressions in the Medical Domain: Who Carries the Domain Specific Meaning	
Kristina Kocijan, Krešimir Šojat and Silva Kurolt	26
Named Entity Recognition with NooJ	
Aleksandar Petrovski	28
Optimization of Portuguese Named Entity Recognition and Classification by Combining Local Grammars and Conditional Random Fields Trained with Parsed Corpora	
Diego Válio Antunes Alves, Božo Bekavac and Marko Tadić	30
Automatic Recognition and Translation of Tunisian Dialect Named Entities into Modern Standard Arabic	
Roua Torjmen and Kais Haddar	32
3. TEXT MINING AND QUESTION ANSWERING.....	35
Annotation of Cause - Result Questions in Standard Arabic Using Syntactic Grammars	
Bessaies Essia, Mesfar Slim and Ben Ghazela Henda	36
A Bottom-Up Approach for Moroccan Legal Ontology Learning from Arabic Texts	
Kautar Belhoucine, Mohammed Mouchid, Aziz Mouloudi and Samir Mbarki	38
Answering Arabic Complex Question	
Sondes Dardour, Héra Fehri and Kais Haddar	40
Development of a Question-Answering System in the Legal Field Based on Ontologies	
Ismahane Kourtin, Aziz Mouloudi and Samir Mbarki	42
4. AI, LEXICONS AND DICTIONARIES.....	45
NooJ for Artificial Intelligence: an Anthropic Approach	
Mario Monteleone	46

New Global Cloud Solutions and NooJ's Woven Digital Intelligences for Homologated Synthetic Fixed Communication	
Raffaele Marcone, Rosa Giulia, Roberto Capone, Giulia Savarese, Marianna Greco, Colomba La Ragione and Janvier Julian Enriquez	48
Standardization and Implementation of Lexicon-Grammar Tables in NooJ Platform	
Asmaa Kourtin, Mohammed Murchid, Abdelaaziz Mouloudi and Samir Mbarki	50
Using Linguistic Software NooJ in Describing Preschool and Younger School Children's Croatian Language Vocabulary	
Katarina Aladrović Slovaček	52
New Russian Resources for Silberztein's Software NooJ	
Vincent Bénet	54
Al-Erfan-DIC: The Electronic Dictionary of Standard Arabic Using NooJ Platform	
Moahmed El Hannach and Ahmed Bounoua	56
Lexical Complexity and Basic Vocabulary of the Italian Language	
Annibale Elia, Alessandro Maisto, Lorenza Melillo and Serena Pelosi	58
Formalising the Latin Language on the Example of Medieval Latin Wills	
Linda Mijić and Anita Bartulović	60
Morphology of Middle French Verbs with NooJ	
Mourad Aouini and Laure-Anne Caraty	62
A Morphological Grammar for Modern Greek: State of Art, Evaluation and Upgrade	
Lena Papadopoulou and Elina Chatjipapa	64
5. SYNTAX AND SEMANTICS	67
Detecting Negation Scope with NooJ	
Gaurish Thakkar, Nives Mikelić Preradović and Jeremy Barnes	68
The Annotation of Indonesian Reduplications with NooJ	
Prihantoro	70
The Automatic Paraphrasing of Arabic Syntactic Structures	
Azeddine Rhazi and Ali Boulaalam	72
Using NooJ for Marketing Choices	
Magaly Bigey	74
Computational Modeling of a Nominal Ellipsis Grammar for Spanish	
Walter Koza and Hazel Barahona	76

6. TEACHING WITH NOOJ	79
Automatic Treatment of Causal, Consecutive and Counterargumentative Discourse Connectors in Spanish: a Pedagogical Application of NooJ	
Andrea Rodrigo, Silvia Reyes and María Andrea Fernández Gallino	80
Preparing the NooJ German Module for the Analysis of a Learner Spoken Corpus	
Mirela Landsman Vinković and Kristina Kocijan	82
7. DIFFERENT SHADES OF VERBS	85
Formalizing Ukrainian Verbs with NooJ	
Olena Saint-Joanis	86
Transformations and Paraphrases for QU Sentiment Predicates	
Maximiliano Duran	88
Verbal syntax in Romani: Diatheses	
Masako Watabe	90
Arabic Psychological Verbs Recognition through NooJ Transformational Grammars	
Asmaa Amzali, Mohammed Mourchid, Abdelaziz Mouloudi and Samir Mbarki	92
Formal Processing of Values of the Simple Kabyle Aorist Using NooJ	
Hamid Annouz	94
Formalizing Arabic Deverbal Noun ‘Gerund’ Using NooJ Platform	
Ilham Blanchete, Mohammed Mourchid, Samir Mbarki and Abdelaziz Mouloudi	96
List of Contributors	99
Index	101

Invited Talks

Digital Dictionary Database and ELEXIS Dictionary Matrix

Simon
Krek

Abstract

In the talk, I will present the idea and the current state of Digital Dictionary Database (for Slovene) which combines traditional lexicographic information such as sense division for individual lemmas or multiword expressions, definitions, usage labels, corpus attestations, collocations, and similar, with NLP-oriented information such as morphological paradigms with patterns, distributional (syntactic) information, semantic roles or frames, semantic types etc. The ultimate goal is to create an extensive collection of language data for a particular language, in this case Slovene, organised in a (semantically-oriented) data model that enables open access to the data in a central online platform.

The system could be defined as Lexicographic Data as a Service (LDaaS) making data available through a REST API considered as part of a general (national) digital infrastructure, comparable with other types of emerging (open (real time)) data used in Artificial Intelligence systems, such as traffic data, street maps, satellite image data etc. An integral part of the system is automatic extraction of linguistic data from monitor corpora (cf. Gantar et al. 2016) in real time that allows constant updating of information in the Digital Dictionary Database. An example of such a procedure is represented by Collocations Dictionary of Modern Slovene (Kosem et al 2018).

Monolingual perspective of collecting and organising extensive linguistic data for a particular language is superseded by a cross-lingual perspective in ELEXIS (European Lexicographic Infrastructure) project (Krek et al. 2018). An important goal of the project is identification of key structural elements found in different types of existing dictionaries, and establishing links between them. As such, it is focused on (direct or indirect) linking of existing lexicographic resources on the sense level by pivoting through one of the existing semantic resources (BabelNet).



Jožef Stefan Institute &
University of Ljubljana
[Ljubljana, Slovenia]



simon.krek@ijs.si



Lexicography
Lexical database
Data modeling
(Lexicographic) Linked
data
Extraction of corpus data
Slovene

2

References

Gantar, Polona, Kosem, Iztok, Krek, Simon. **Discovering automated lexicography: the case of Slovene lexical database.** International journal of lexicography, 2016, vol. 29, issue 2, pp. 200-225.

Kosem, Iztok, Krek, Simon, Gantar, Polona, Arhar Holdt, Špela, Čibej, Jaka, & Laskowski, Cyprian. **Collocations dictionary of modern Slovene.** Proceedings of the 18th EURALEX International Congress. Ljubljana: Ljubljana University

Press, Faculty of Arts. 2018, pp. 989-997.

Krek, Simon, Kosem, Iztok, Mccrae, John P., Navigli, Roberto, Pedersen, Bolette S., Tiberius, Carole, Wissik, Tanja.

European Lexicographic Infrastructure (ELEXIS). In: ČIBEJ, Jaka (ed.), et al. Proceedings of the 18th EURALEX International Congress. Ljubljana: Ljubljana University Press, Faculty of Arts. 2018, pp. 881-891.

The aim of this activity is to produce a massive semantic “dictionary matrix” leading to the possibility of establishing a lexicographically-based sense or concept repository useful for Natural Language Processing and Artificial Intelligence. Following the idea of Universal Dependencies project (together with its universal POS tag set) one can imagine the emergence of “Universal Concepts”, empirically verifiable through links available in the ELEXIS dictionary matrix.

In the talk, the current developments in the project will be presented with a special focus on (dictionary) sense linking and word sense disambiguation tasks, in the context of a shared task on the task of monolingual word sense alignment across dictionaries organised by ELEXIS as part of the GLOBALEX 2020 workshop, and the work on a manually annotated corpus for 10 languages, used for cross-lingual word sense disambiguation.

Data to Help us Form Images: the Abstract and Non-Imageable

Anita
Peti-Stantić

Abstract

When we say ‘a chair’ or ‘a dog’, we instantly and unconsciously imagine the picture of (a) dog or (a) chair. This does not happen when we say ‘peace’ or ‘experience’.

Since numerous psycholinguistic and neurolinguistic studies, as well as educational models, use words in order to measure latencies and the accuracy of word recognition, they have to somehow account for this difference in their properties. Core concepts laying behind these properties are concreteness and imageability. They are claimed to play a significant role in remembering, recognizing and understanding not only words, but the complexity of the language architecture as well. Since normed ratings are not readily available for understudied languages such as Croatian, I will present *Croatian Psycholinguistic Database* freely available at: <https://doi.org/10.17234/megah.2019.hpb>.

This normative database was constructed within the project funded by the Croatian National Foundation: *The Building Blocks of Croatian Mental Grammar: Constraints of Information Structure* (Peti-Stantić et al. 2018). The database contains empirically established norms for the categories of concreteness, imageability, subjective frequency and age of acquisition for 6000 lexemes of Croatian words excerpted from the hrWaC web corpus of Croatian (Ljubešić & Klubička, 2016) by combining the hrLex inflectional lexicon with objective word frequencies from hrWac. The objective characteristics of each word are also coded, i.e., word length in number of characters, word class (noun, verb, adjective, and adverb), animacy and gender for nouns, and raw frequency in the hrWaC corpus.

The ratings were collected in two experiments with 3,000 words tested in each (Peti-Stantić et al. in print). A total of 3,630 questionnaires were completed by the native speakers of Croatian, while every word was



Faculty of Humanities
and Social Sciences
[Zagreb, Croatia]



anita.peti-stantic
@ffzg.hr



*Croatian psycholinguistic
database*

Concreteness

Imageability

Word class

Croatian language

4

References

Ljubešić, N. & Klubička, F. (2016). **Croatian web corpus hrWaC 2.1**, Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1064>.

Peti Stantić, A., Anđel, M., Keresteš, G., Ljubešić, N., Stanojević, M.-M., & Tonković, M. (2018). **Psycholinguistic estimates of 3000 words of Croatian: Concreteness and imageability**. *Suvremena lingvistika*, 44(85), 91–112. <https://doi.org/10.22210/suvlin.2018.085.05>

Peti-Stantić A., Anđel M. Gnjiđić V., Keresteš, G., Ljubešić N., Masnikosa I., Tonković, M., Tušek, J., Willer-Gold, J., Stanojević, M.-M. (in print) **The Croatian Psycholinguistic Database: Estimates for 6,000 Nouns, Verbs, Adjectives and Adverbs.** *Behavior Research Methods.*

rated by an average of 30 participants. Analyses replicate the correlations reported in the literature, whereby words rated as more concrete are more highly imageable, shorter, acquired earlier and rated as more subjectively frequent. Differences in concreteness, imageability, subjective frequency and age of acquisition are reported amongst nouns, verbs, adjectives and adverbs.

I will discuss the data collected and differences between word classes from the perspective of constructional cognitive linguistic theories. I will also show the applicability of the database for researchers and practitioners.

Digital Humanities

NooJ

for the Digital Humanities

Max
Silberstein

Abstract

Most software applications used in the Digital Humanities by researchers in Social Sciences (History, Literature studies, Psychology, Sociology) analyze texts by processing occurrences of simple orthographic word forms, and then analyzing their frequency (Dacos&Mounier, 2015). However, linguists have known for a long time that there is no direct correspondence between orthographic word forms and the concepts and entities that are relevant for the researchers in the Social Sciences. In consequence, most results produced by these software applications are unreliable at best.

We propose to present a new software application based on the NooJ linguistic engine (Silberstein, 2015), “NooJ for the Digital Humanities”, that is specifically aimed at bridging the gap between what the software processes (orthographical word forms) and what users need (concepts, entities and relations that can be expressed as linguistic units). The new software application manages orthography so that its statistical analyzer counts together occurrences of variants such as “audiovisual” and “audio visual”; it manages inflection and derivation so that its statistical analyzer counts together all forms of a concept such as demonstration and demonstrators.

The new software application can also process lexical fields defined as lists of related words and expressions, such as:

DEATH = <death> | <kill> | <perish> | <cemetery> | funerals
| <murder> | ...

LOVE = <love> | <lover> | <adore> | <affection> | <friendship>
| <tenderness> | ...



Université de Franche-Comté
[Besançon, France]



max.silberstein
@univ-fcomte.fr



Digital humanities
Statistical analysis
Corpus linguistics
Information retrieval
NooJ

8

References

Dacos M., Mounier P., 2015. **Humanités Numériques : État des lieux et positionnement de la recherche française dans le contexte international**. Rapport de 8e recherche: <http://hal.archives-ouvertes.fr/hal-01228945>.

Silberstein M., 2015. **Formalising Natural Languages: The NooJ Approach**. Wiley: Hoboken.

Silberstein M., 2018. “Using linguistic resources to evaluate the quality of annotated corpora”. In Proceedings of the LR4NLP Workshop at

COLING2018:
<http://www.aclweb.org/anthology/W18-38>.

By processing these lexical fields as units, users can then perform narrative analyses of their corpus to detect if a novel contains a love story, to see if it ends badly, to compare all novels from an author, etc.

Lexical fields might contain syntactic description in order to remove non-relevant homographs. For instance, the French noun “tombe” stands for “tomb” and would be relevant to the MORT [DEATH] lexical field. However, the word form “tombe” can also represent a conjugated form of the verb “tomber” [to fall]. In order to get only occurrences of the noun, it is possible to insert syntactic context in the lexical field definition, e.g. “<DET> <A>* tombe”. In consequence, the statistical analyzer will count the word form “tombe” as an occurrence of the lexical field MORT [DEATH] only when it occurs in a syntactic context of a Noun.

The software application offers its users all standard statistical measurements, such as evolution of the vocabulary, specificity of a text, relevance of a concept, multiple correspondence factorial analysis, etc. Finally, we evaluate the new approach by comparing results of statistical analyses computed on word forms with those computed on linguistic units (Silberztein, 2018).

NooJ in the Humanities: Phrasal Verb Usage in the Works of British and American Authors

Peter A.

Machonis

Abstract

Since Kennedy's (1920) seminal study, phrasal verbs (PV) have often been classified as pleonastic or colloquial variants of simple verbs (*finish up vs. finish; cough up vs. pay*). Linguists, however, have pointed out a steady growth of PV after Old English, a slight drop during the Age of Reason, followed by a new expansion in the 19th century. Today though, PV are often attributed to an American influence by grammarians. For example, The New Fowler's Modern English Usage (2000:594) states: "Frequent in American English, it is clear that the use of phrasal verbs began to increase in a noticeable manner in America from the early 19th century onward. From there, many have made their way to Britain during the 20th century, to widespread expressions of regret and alarm."

However, Stephan Thim (2012:203-5) claims "little evidence for the universal assumption that phrasal verbs are more typical of American English." In fact, he highlights "the little attention Late Modern English – in particular the 19th century – has received." This study thus proposes to expand digital research in the humanities by using NooJ, along with a specially designed PV grammar, electronic dictionary, and a series of disambiguation grammars, adverbial and adjectival expression filters, and idiom dictionaries to parse the complete works of Charles Dickens and his American counterpart, Herman Melville and compare their PV usage. Although digital humanities date back to the first concordances, this project is unique in that it involves identifying a precise type of verb which can appear in many contexts.

In fact, automatic recognition of English PV is more complex than for other multi-word expressions, due to their possible discontinuous nature, their confusion with verbs followed by simple prepositions, and genuine ambiguity – only resolvable from context. Even with



Florida International
University
[Miami, Florida, USA]



machonis@fiu.edu



English phrasal verbs
Corpora linguistics
Charles Dickens
Herman Melville
NooJ

References

Burchfield, Robert William. 2000. **The New Fowler's Modern English Usage**. Rev. 3rd ed. Oxford; New York: Oxford University Press.

Kennedy, Arthur Garfield. 1920. **The Modern English Verb-Adverb Combination**. Stanford: Stanford University Press.

Thim, Stephan. 2012. **Phrasal Verbs: The English Verb-Particle Construction and Its History**. Berlin: Walter de Gruyter.

disambiguating grammars and other filters, our previous studies achieved only 88% accuracy, with most of the noise coming from the particles *in* and *on*, which are fairly complicated to distinguish automatically from prepositions (e.g. the prepositional phrase *had a strange smile on her thin lips* vs the PV *had her hat and jacket on*).

Consequently, for this corpora study on the novels of Dickens and Melville, we limited PV searches to six typical particles representing three levels of PV frequency: high (*out, up*), mid (*down, away*), and low (*back, off*). In fact, if we limit searches to include only these six particles, we can achieve 98% accuracy, including identifying many discontinuous PV in both Dickens (*I still held her forcibly down; the moment she appeared, Joe took his hat off and stood; If you bring the boy back with his head blown to bits by a musket*) and Melville (*Canst thou not drive that old Adam away?; The sea had jeeringly kept his finite body up, but drowned the infinite of his soul; he had that club-hammer there ... to knock some one's brains out with, I suppose*).

Since usage could be attributed to subject matter, we analyzed usage per 1,000 words of text in the complete works of Melville (1.3 million words) and Dickens (4 million words), obtained from Project Gutenberg. Our preliminary results show that Dickens uses more PV than Melville: 3.39 PV per 1,000 words of text as compared to 2.5 PV per 1,000 words of text.

To avoid an author bias, future research will examine more American and British authors from this time period, showing to what extent an American influence, if any, was involved in the expansion of PV. Furthermore, our PV program would play a central role in using digitalized resources in the humanities – connecting the past to the present and the future of research.

Maria Deraismes' Writings on Women: Study Using NooJ Linguistic Platform


Odile
Piton

Abstract

We propose to carry out a study of texts by Maria Deraismes (1828-1894). Through her actions, writings, and lectures, she played an important part in the history of feminism in France. Endowed with a political sense, she was a republican, anti-clerical, and a free-thinker. Born into a well-to-do family, she benefited from an exceptional education, and acquired encyclopedic knowledge, great culture, and a vast scholarship. She frequented intellectual circles. When asked to give lectures, her talent as an orator aroused enthusiasm. She became famous for her writings as a journalist, then as director of the newspaper "*Le Républicain de Seine et Oise*", and as a lecturer, and created the Society for the Improvement of the Status of Women and the Defense of Women's Rights.

She was engaged in a critical study of 19th century French society, analyzing its dysfunctions and in particular the civil and political subordination of women. She affirmed: "The inferiority of women is not a fact of nature, we repeat, it is a human invention, it is a social concept." She denounced the "narrow and erroneous" education reserved for women. She advocated working for the reform of society. She invited thinking beyond the social norms of the time, analyzed the role that was assigned to women, and suggested that they should combine their talents, cultivate collective activity, and abandon the reading of novels in favor of serious readings: "women have the right and the obligation to take an interest in public matters". She strongly criticized the influence of the Church and the subdued role it assigned for women: "women have been held longer than men under the yoke of superstition and ignorance", and she turned to Freemasonry. She set up the lodge *Le Droit Humain* an equal men and women obedience.

 SAMM EA 4543 / CNRS
FR2036,
Université Paris1
Panthéon-Sorbonne
[Paris, France]

 piton@univ-paris1.fr

 NLP
19th century
Feminism
Maria Deraismes
French language
NooJ

References

Boime A. (1994) **Maria Deraismes and Eva Gonzalès: A Feminist Critique of 'Une Loge Aux Théâtre Des Italiens**. *Woman's Art Journal*, vol. 15, no. 2, 1994

Pignot H., Piton O., (2011) **Mary Astell's words in A Serious Proposal to the Ladies (part I), a lexicographic inquiry with NooJ**, Proceedings of the NooJ 2010 International Conference and Workshop. Publication of the

Democritus University of Thrace,
Komotini, 2011. 232-44.

Singer Cl. (dir.) (2012) **Maria
Deraismes, journaliste
pontoisienne : une féministe et
libre-penseuse au XIX^{ème}
siècle**, Karthala ed., coll.
Tropiques.

We selected some of her writings concerning women and their legal or social status in France. Her texts include many Latin quotations and political, historical, and sociological references, particularly on the role of women and female deities in history.

The French language presents some modifications since the XIXth century. These variations do not bother the reader but they are obstacles to automatic processing. We show the help provided by NooJ in dealing with the morphological, derivational, and lexical particularities of these texts. It provides us with adequate tools to carry out a linguistic and pragmatic study of these texts. We denote that some of Deraismes' observations are in line with current preoccupations.

Depictions of Women in “Duga” and “Tena” - a Computational Analysis

Lorena
Kasunić

Gordana
Kiseljak

Abstract

By combining two contrasting fields, the humanities and digital technologies, a new discipline was born – that of digital humanities. After a long period of constantly using a single approach in literary science – interpretation – in recent years, a new approach has been introduced in this field of science – an empirical type of research, that of computational analysis. The most prominent work using computational analysis for literary text analysis has been done (and is still being done) at institutions such as the Stanford Literary Lab. Not many efforts have been made in Croatia to implement or even introduce computational analysis as a possible approach in literary studies.

This paper analyzes depiction of women in two particular short stories: “Tena” by Josip Kozarac and “Duga” by Dinko Šimunović. These texts are considered appropriate and representative because their main characters are women whose portrayals are given in great detail. That is the reason why analysing such texts can be fruitful for deriving data for the construction of a general model. Both texts are written in the Croatian language, which points to one of the main purposes and intentions of this paper - using canonical literary texts from Croatian literature to build a model for quantitative analysis of female characters and hopefully apply it to other texts in the future. An example of a classification of women can be found in the work of Božidar Petrač, a Croatian literary historian, who provides the following archetypes (1990): woman as a mother, woman as an ideal woman, woman as a hero, woman as someone unreachable, woman as a lady, woman as a sinner, the penitent woman, woman as a “femme fatale”. Computational text analysis gives us the opportunity to explore more about women as characters than the aforementioned labels.



Faculty of Humanities
and Social Sciences
[Zagreb, Croatia]



lkasunic@ffzg.hr
gkiselja@ffzg.hr



Computational analysis
Depictions of women
Digital humanities
Croatian language
NooJ

14

References

Bamman D., Smith N.A., Massey, P. and Xia, P. (2015). **Annotating Character Relationships in Literary Texts**. [online] arXiv. Available at: <https://arxiv.org/abs/1512.00728>. [Accessed: 28 Apr. 2020].

Bullard, J. and Ovesdotter Alm, C. (2014). **Computational analysis to explore authors' depiction of characters**. In: Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL). [online] Gothenburg: Association for Computational Linguistics, pp. 11-16. Available at:

<https://www.aclweb.org/anthology/W14-0902.pdf>. [Accessed: 28 Apr. 2020].

Jacobs, A.M. (2019). **Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics**. [online] *frontiers*. Available at: <https://www.frontiersin.org/articles/10.3389/frobt.2019.00053/full> [Accessed: 28 Apr. 2020].

Petrač, B. (1990). Lik žene u hrvatskoj književnosti, *Bogoslovska smotra*, [online] Volume 60(3-4), pp. 348-354. Available at: <https://hrcak.srce.hr/37292> [Accessed: 28 Apr. 2020].

A study conducted by Bullard and Ovesdotter Alm (2014) is an example of how depictions of characters can be explored by computationally analysing character's written speech. Jacobs (2019) is also engaged in character building, but his research provides emotional and personality profiles using the principles of computational poetics. Similarly, Bamman, Smith, Massey et al. (2015) investigate character relationships by manually annotating them. This is part of the training and evaluation of a model which automatically predicts types of relations between characters.

The methods which are used in this paper are computational analysis (quantitative approach) and interpretation (traditional, qualitative approach) of quantitative results based on literary science. This paper represents a pilot study of sorts, in which the authors will use the NooJ environment to make the initial steps toward building the aforementioned model. The authors maintain that enriching the traditional approach with empirical results can lead to new possibilities in the interpretation of literary texts.

The Use of Figurative Language in a Dream Descriptions Corpus. Exploiting NooJ for Stylometric Purposes

Raffaele
Manna

Antonio
Pascucci

Maria Pia
di Buono

Johanna
Monti

Abstract

Computational Stylometry (CS) develops techniques that allow scholars to find out information about authors of texts by means of an automatic stylistic analysis, which aims at identifying/profiling an author. Indeed, each author's style is unique and there are no authors characterized by the same set of stylistic features, which constitute the so-called fingerprint and authorial DNA.

In order to represent the fingerprint and authorial DNA, several scholars focus on the analysis of different stylistic features and specific linguistic phenomena. Still, complex semantic aspects of figurative language remain one of the less investigated stylistic features (Benotto, Giovannetti, and Marchi, 2016), due to the lack of a precise mapping between linguistic realizations and concepts underlying the figurative language.

Since metaphors, similes, metonymies, and the whole figurative language play an important role in how we mold the day-to-day reality, they are represented through specific linguistic phenomena suitable to express specific conceptualization. Indeed, according to the Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 2008), Linguistic Metaphors (LMs) represent a part of a broader generalization known as Conceptual Metaphors (CMs).

Several linguistic devices and markers may highlight the presence of metaphorical expressions and other figures of speech, e.g., metalinguistic expressions such as metaphorically/figuratively speaking or so to speak, general metalanguage about semantics such as in more than one sense, 'mimetic terms' like image, likeness or picture,



L'Orientale University of
Naples - UNIOR NLP
Research Group
[Naples, Italy]



rmanna@unior.it
apascucci@unior.it
mpdibuono@unior.it
jmonti@unior.it



*Computational
stylometry
Figurative language
processing
Distributional semantics
Corpus analysis
Italian language
NooJ*

References

Benotto, G., Giovannetti, E., & Marchi, S. (2016). **Investigating the Application of Distributional Semantics to Stylometry**. *CLiC it*, 61.

Bruni, E., Tran, N. K., & Baroni, M. (2014). **Multimodal distributional semantics**. *Journal*

of Artificial Intelligence Research, 49, 1-47.

Goatly, A. (1997). **The language of metaphors**. Routledge.

Lakoff, G., & Johnson, M. (2008). **Metaphors we live by**. University of Chicago press.

Manna, R., Pascucci, A., & Monti, J. (2019). **Gender Detection and Stylistic Differences and Similarities between Males and Females in a Dream Tales Blog**. In CLiC-it 2019 Sixth Italian Conference on Computational Linguistics (Vol. 2481).

intensifiers like literally, actually, veritable, etc., and even orthographic devices like quotation marks (Goatly, 1997).

Our goal is the exploitation of NooJ functionalities to

- i) build a domain dictionary able to represent lexical and semantic characteristics of metaphorical expressions;
- ii) model grammars to recognize, extract, and tag these metaphors and figures of speech, on the basis of different linguistic devices.

The proposed methodology for automatic recognition of metaphors relies on the analysis of semantic distributions of these phenomena. In fact, distributional semantics, namely a series of computational methods for the study of semantic distribution in texts, allows representing word meaning from the patterns of co-occurrence of words in a text (Bruni, Tran, and Baroni, 2014).

The research has been carried out using a corpus (Manna, Pascucci, and Monti 2019), which collects Italian blog posts about dream descriptions from several users.

Paraphrasing Emotions in Portuguese

Cristina
Mota

Diana
Santos²

Anabela
Barreiro

Abstract

Emotions are often referred to with language and, in fact, according to some authors (Wierzbicka 1999), they show the highest variability in forms of expression.

For example, in Portuguese, the following expressions:

- *ele apaixonou-se por*
- *ele estava apaixonadíssimo por*
- *ele é um apaixonado por*
- *ele ficou apaixonado por*
- *ele teve uma paixão por*

are all syntactic variations of the same predicate in verbal, adjectival or nominal form. Although there are subtle and not so subtle semantic differences between these examples, they clearly show that emotions offer wide paraphrasability means.

By analysing five different texts in two varieties of Portuguese, namely Portuguese from Portugal and from Brazil, corresponding to more than 2,500 aligned sentences, this paper intends to find cases of emotion paraphrases.

We use the emotion framework devised in Mota and Santos (2015) – developed to annotate the AC/DC corpora – and revise the results. The emotion paraphrases will then be aligned with CLUE-Aligner.

The second step is the enrichment of the paraphrastic resources of Port4NooJ, the Portuguese module of NooJ, with emotion information. This will be done by merging this new information with the one previously derived from lexicon grammar tables of human intransitive adjectives and constructions with the support verbs *fazer* and *ser de* (cf., for instance, Mota, Baptista, and Barreiro (2019)).



INESC-ID & Linguatca
[Lisbon, Portugal]

²Universidade de Oslo &
Linguatca
[Oslo, Norway]

cmota
@ist.utl.pt



d.s.m.santos
@ilos.uio.no

anabela.barreiro
@inesc-id.pt



*Emotions
Paraphrasing
Polarity lexicon
AC/DC corpora
Portuguese language
NooJ*

References

Mota, Cristina, Jorge Baptista, and Anabela Barreiro (2019). **The Lexicon-Grammar of Predicate Nouns with ser de in Port4NooJ.** In *12th International Conference, NooJ 2018, Palermo, Italy, June 20–22, 2018, Revised Selected Papers*, pp. 124–137.

Mota, Cristina and Diana Santos (2015). ***Emotions in natural language: a broad-coverage perspective***. Tech. rep. Linguateca.

Wierzbicka, Anna (1999). ***Emotions across languages and cultures: Diversity and universals***. Cambridge University Press.

Finally, we will use Port4NooJ within the eSPERTo paraphrasing system to suggest paraphrase of emotions and evaluate the results.

Text Analysis of Scientific Papers using NooJ: Case Study with Vitamin D Supplementation Debate

Lesia

Kaigorodova

Abstract

This work is dedicated to text mining of scientific papers with NooJ regarding vitamin D consumption, probably the most popular topic nowadays in the era of supplements.

We will try to design a system that would provide answers to such questions as:

- do we benefit from vitamin D intake or is it harmful, and what are the restrictions?
- how does supplementation work in the presence of a particular disease?
- can we provide medical practitioners with up-to-date knowledge in the domain?

The research can be extended to address other queries that may refer to medicine as well.

The data for the research was collected from the PubMed database of the NCBI (United States National Center for Biotechnology Information) website. It has the greatest selection of articles in the field of medical domain. Overall 83 077 articles were scraped from the PubMed database that refer to the keyword “vitamin D”. The data is comprised of a title, list of authors, date, journal, citation and an abstract. Some abstracts contain more structured information in the form of sections such as “conclusion”, “objectives”, “method”, “results”, “design”, “setting”.

Automatic Opinion Analysis is used in order to address the debate that has occurred in the modern world regarding vitamin D consumption. It



UIIP, National Academy of Sciences of Belarus [Minsk, Belarus]



lesia.piatrouskaya@gmail.com



Automatic opinion analysis
NLP
Text mining
Vitamin D
Medical domain
Belarus language
NooJ

References

Hetsevich, Yu. **Grammars for Sentence into Phrase Segmentation: Punctuation Level**. In *Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015, Minsk, Belarus, June 11-13, 2015, Revised Selected Papers* / ed. T. Okrut, Y. Hetsevich, M. Silberstein, H. Stanislavenka. — Springer

International Publishing, 2016.
pp. 74-82.

Silberztein Max, 2018. **NooJ Manual**. Available for download at: <http://www.nooj-association.org>.

should be mentioned that specific terminology and language patterns should be taken into consideration in order to process scientific papers.

The concept for "vitamin D" is introduced as well as positive and negative features for the outcome of vitamin D supplementation. For example, the "vitamin D" concept comprises {vitamin D, cholecalciferol, D2, D3, 1,25-dihydroxyvitamin, 25(OH), VD, ...}. For "positive features" of the outcome it may be {promising, good, help, safe, improve, stabilize, regulate, positive, progress, ...}. For the "negative features" it may be {negative, dangerous, unsafe, side effect, cautious, abnormal, harmful, trigger, worse, ...}. It should be noted that positive and negative features should be located and processed in the text according to its context since it can be negated and hence represent the opposite.

The grammars are constructed to make all the introduced concepts work together in order to extract meaning from the abstracts, i.e. positive, negative or neutral outcome of the vitamin D intake. Other additional information such as sections "objectives", "conclusion", etc. is analyzed for the inference of particular knowledge.

With NooJ it may be possible to design a knowledge-based system for the available information in the domain of medicine. It could be helpful for both medical practitioners to make information-driven decisions and for individuals who are desperate to find the truth in the abundance of contradicting points of view.

Automatic Extraction of French Foods Expression Routine with *NooJ*

Tong
Yang

Abstract

Our study fits the teaching method FOS (Mangiante and Parpette, 2004) for Chinese cooks who come to work in French restaurants. The routines that constitute a category for discourse analysis make it possible to account for the association of form and function.

According to Tutin & Kraif (2016), routines have four characteristics:

1. They are linear, sometimes with "holes" in the sequences;
2. They include sequences of words, lemmas or morphosyntactic features;
3. They do not necessarily constitute classical syntactic constituents;
4. They are characterized by a specific discursive, textual, or rhetorical function.

In the culinary field, the French foods expression routine *objet + à + objet* are very recurrent (e.g., *tarte à la crème*, *roulade de saumon fumé aux asperges*, *sauce aux crevettes*). Before teaching these routines to our students, the extraction of this structure becomes an unavoidable task.

However, the extraction of sequences of multiple words always poses a problem in the TAL (automatic language processing) (Luka et al., 2006). According to the needs (disambiguation and flexibility) of our extraction, NooJ becomes the most appropriate software, because in theory, NooJ can describe all the natural languages of the world (Silberztein, 2015; 2016).

The extraction of these routines with NooJ is therefore the main problem encountered in our research in which we will present, first of all, our object of study: French foods expression routine. Then, the



Université de Paris 3 &
Université de Franche-
Comté
[Paris & Besançon,
France]



tong.yang
@sorbonne-nouvelle.fr



Automatic extraction
French food
Expression routine
Disambiguation
Cuisitext
French language
NooJ

References

Mangiante, J.M. & Parpette, C., (2004). **Le français sur objectif spécifique : de l'analyse de besoins à l'élaboration d'un cours**. Hachette.

Silberztein, M., (2015). **La formalisation des langues: l'approche de NooJ**, International Society for Technology in Education.

Yang, T., (2017). **Automatic Extraction of the Phraseology**

Through NooJ. In: Mbarki S., Mouchid M., Silberstein M. (eds) Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications. NooJ 2017. Communications in Computer and Information Science, vol 811, pp.167-177, Springer, Cham.

modeling and the disambiguation will be highlighted to achieve the automatic extraction. Inspired by our previous projects (Yang, 2017; 2018; 2019), our modeling is based on observations found in Cuisitext (Yang, 2016) and our disambiguation concerns both the nouns and adjectives of recognized expressions. Take the rejected names for example, <ce>, <la>, <être>, <avoir>, <bien>, <si>, <été>, <y>, <g>, <h>, <tout>.

Finally, we will implement the data in NooJ by elaborating some grammars in the form of a transducer and compile our data according to the NooJ codes, in particular the variable (prefixed by the character "\$"), the global variable (prefixed by the character "@") and the colored nodes (prefixed by the character ":").

Multi Word Expressions and Named Entites

Multiword Expressions in the Medical Domain: Who Carries the Domain Specific Meaning

Kristina
Kocijan

Krešimir
Šojat

Silvia
Kurolt

Abstract

This paper is a continuation of work in natural language processing in the medical domain for Croatian (Kocijan et al., 2019). After we have annotated the single nouns from the corpus consisting of pharmaceutical instructions for medicaments (Kocijan et al., 2020), we are ready to shift the focus to multiword expressions. The project will still rely on the nouns from the previous step to detect MWEs where the noun is the main carrier of the medical meaning (e.g. acute **pain** – ‘*akutna bol*’).

However, in the cases where the main noun is more general and not directly associated with the medical domain e.g.

- kidney **function** - ‘*bubrežna funkcija*’;
- pharmaceutical **data** - ‘*farmaceutski podatci*’;
- endoscopic **results** - endoskopski **nalazi**’;
- histological changes – ‘*histološke promjene*’;
- gastrointestinal system - ‘*gastrointestinalnom sustavu*’,

we will use the power of NooJ morphology grammar to check if the preceding adjective root is associated with the noun found in the main dictionary and annotated as a medical domain noun.

Thus, we are checking if the adjective endoscopic (*endoskopski*), has a noun endoscopy (*endoskopija*) that is in NooJ dictionary already marked as a noun belonging to the medical domain. In such cases, we will assume that the adjective belongs to the same domain and that the attribute for the medical domain can be inherited, not only for the adjective, but for the entire MWE as well.



Faculty of Humanities
and Social Sciences
[Zagreb, Croatia]



krkocijan@ffzg.hr
ksojat@ffzg.hr
skurolt@ffzg.hr



MWE
Medical domain
adjectives
Croatian language
NooJ

References

Kristina Kocijan, Maria Pia di Buono, Linda Mijić: **Detecting Latin-based Medical Terminology in Croatian Texts**, In (I. Mauro Mirto; M. Monteleone; M. Silberstein; eds.) *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*. Communication in Computer and Information Science, 987. Springer International Publishing, pp. 38-49, 2019.

Kristina Kocijan, Silvia Kurolt, Linda Mijić: **Building Croatian**

**Medical Dictionary from
Medical Corpus**, In *Rasprave:
Časopis Instituta za hrvatski jezik
i jezikoslovlje* (2, 46) 2020.

The project hopes to help with automatic extraction and annotation of single adjectives from the medical domain, but also to help identify medical MWEs. The grammar results will be analyzed to check for the precision and recall in order to make necessary augmentations.

However, we are also hoping to learn more about who carries the domain specific meaning in Croatian MWEs.

Named Entity Recognition with NooJ

Aleksandar
Petrovski

Abstract

The term “named entity” (NE) refers to expressions describing real-world objects, like persons, locations, and organizations. Named entities are often denoted by proper names. They can be abstract or have a physical existence. Some other expressions, describing money, percentage, time, and date might also be considered as named entities. Examples of named entities include Great Britain, Zagreb, Google, Manchester United, Microsoft Windows, September 11th, or anything else that can be named.

The role of named entities has become more and more important in Natural Language Processing (NLP). Named entity recognition (NER) is an NLP task that automatically detects named entities in texts. It was first introduced to the NLP community at the end of the 20th century. Named entity recognition is crucial in information extraction. Some use cases of NER technology include: machine translation, content classification, improving the efficiency of search algorithms, powering content recommendations, customer support, etc.

This paper deals with NER in Macedonian texts, using NooJ. For that purpose, several lexical resources have been used:

- a huge general morphological lexicon,
- morphological lexicons of proper nouns and toponyms, and
- a lexicon of named entities extracted from Wikipedia.

Using the morphological lexicons, NooJ is capable of recognizing inflectional forms of words. The lexicon of named entities extracted from Wikipedia is comprised of only their canonical forms, which means that their inflectional forms will not be recognized. A mitigating factor



International Slavic
University
[Sv. Nikole, North
Macedonia]



a.petrovski.sise
@gmail.com



*Named entity
Macedonian language
NooJ*

References

Petrovski, A. “**A Morphological Computational Lexicon – a Contribution to the Macedonian Language Resources**”, Ph.D. thesis, University St. Cyril and Methodius, Faculty of natural sciences and mathematics, Institute of informatics, Skopje 2008

Petrovski, A. “**A Computer Dictionary for Toponyms**”, Macedonian language international gathering in the event “2008 - Year of the Macedonian language”, Ohrid 2008

here is that named entities, unlike common nouns, are not very inflective in Macedonian.

In addition to lexical resources, several morphological and syntactic grammars have been created to enhance the results. Here, in accordance with Macedonian orthography, rules related to capitalization have been used.

Syntactic grammars have also been used to recognize expressions describing money, percentage, time, and date. To verify the system, the resources have been applied to a small corpus.

Optimization of Portuguese Named Entity Recognition and Classification by Combining Local Grammars and Conditional Random Fields Trained with Parsed Corpora

Diego
Alves

Božo
Bekavac

Marko
Tadić

Abstract

Named Entity Recognition and Classification (NERC) involves recognizing information units such as person, organization, location (sometimes) and others present inside unstructured texts. For the Portuguese language (both native and Brazilian), the Second HAREM evaluation campaign (Freitas et al. 2010) established a complex, and therefore, ample set of tags for this task and evaluated several NERC systems based on either machine learning approaches or hand-coded rules in combination with dictionaries, gazetteers, and ontologies.

A rule-based model for Portuguese NERC has been proposed inside Port4NooJ v3.0 (Mota et al. 2016), however, the evaluation of this tool has not been provided. Pirovani, J. and De Oliveira, E. proposed in their 2018 article the association of local grammars with Conditional Random Fields (CRF) probabilistic method based on a training set containing Part-of-Speech (PoS) tags in order to enhance NERC results for Portuguese.

Our proposal is to focus on the “Tempo” (time) annotation. First, graphs and additional specialized dictionaries will be created with NooJ in order to pre-annotate the text following HAREM structure and rules for this category (and subcategories). The pre-annotated test text will be, then, processed by a CRF tool trained with a part of HAREM database containing also linguistic information like described above.

Precision, recall and F1 measure will be calculated in each step, as this will allow us to compare the contribution of each method to the final



Faculty of Humanities
and Social Sciences
[Zagreb, Croatia]



dfvalio@ffzg.hr
bbekavac@ffzg.hr
marko.tadic@ffzg.hr



*Named entity
Recognition and
classification*
Local grammars
*Conditional random
fields*
Portuguese language
NooJ

References

Freitas, C., Mota, C., Santos, D.,
Oliveira, H.G., Carvalho, P.,
**Second HAREM: Advancing the
State of the Art of Named Entity
Recognition in Portuguese,**
LREC, 2010.

Mota, C., Carvalho, P., Barreiro, A., **Port4NooJ v3.0: Integrated Linguistic Resources for Portuguese NLP**, LREC, 2016.

Pirovani, J., Oliveira, E., **Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars**, LREC, 2018.

results. CRF method will also be evaluated alone. The peculiarity of our work is that we are using numerous resources from NooJ in order to already have a high precision in the first annotation step, and by adding Parsing annotation as another linguistic feature for the CRF tool. Our hypothesis is that this combination will generate better overall results for NERC in Portuguese, and, therefore, will be able to be expanded to other languages.

Automatic Recognition and Translation of Tunisian Dialect Named Entities into Modern Standard Arabic

Roua
Torjmen

Kais
Haddar



Faculty of Economics and Management of Sfax, Miracl Laboratory [Sfax, Tunisia]



rouatorjmen@gmail.com
kais.haddar@yahoo.fr



*Named entities
Translation
Finite transducer
Tunisian dialect
Arabic language
NooJ*

References

Torjmen R., Ghezaiel Hammouda N., Haddar K.: **A NooJ Tunisian Dialect Translator**. In: Fehri H., Mesfar S., Silberztein M. (eds) Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications. NooJ 2019. Communications in Computer and Information Science, vol 1153. Springer, Cham (2020).

Torjmen R., Haddar K.: **Construction of Morphological**

Abstract

Named entities (NEs) are enormously widespread in different corpora. Tunisian Dialect (TD) corpus is no exception and also contains a huge number of NEs. Moreover, non-Tunisian Arabs find TD difficult to understand. These incentives prompt us to build a tool to recognize NEs in TD corpus and to translate them into Modern Standard Arabic (MSA). Furthermore, this tool offers the possibility of tackling other fields such as information retrieval, automatic indexing, clustering and classification of documents.

Among the problems that may be encountered, TD does not have a standard spelling. This dialect also contains words of different origins such as Arabic, Amazigh, French, Maltese, Spanish, and Turkish. Besides, it differs from one region to another. Furthermore, TD proper nouns can be written in different manners. In addition, there are no capital letters at the beginning of a word to indicate the presence of a TD proper noun. Sharing the same problems as MSA, the absence of diacritical marks and the presence of agglutination phenomenon aggravate the situation and produce homographs which, in turn, can create ambiguities.

The main objective of this paper is to build a tool of recognition and translation of TD named entities into MSA using NooJ linguistic platform. To do this, we start by reusing a bilingual TD-MSA dictionary. Then, we establish a set of syntactic grammars that offers NEs recognition and their translation from TD into MSA. Indeed, our tool is realized using finite state transducers implemented in NooJ linguistic platform. Moreover, we can experiment with our tool thanks to already designed resources (bilingual TD-MSA dictionary, inflectional, derivational and morphological grammars). In addition, we collect texts from social

Grammars for the Tunisian

Dialect. In: Mauro Mirto I.,
Monteleone M., Silberztein M.
(eds) Formalizing Natural
Languages with NooJ 2018 and
Its Natural Language Processing
Applications. NooJ 2018.
Communications in Computer
and Information Science, vol
987, pp. 62-74. Springer, Cham
(2019).

Silberztein, M.: **The
Formalisation of Natural
Languages: The NooJ Approach**,
346 p. Wiley, Hoboken (2016).

networks like Facebook and Twitter to construct two corpora. By detailing, one is for the study and the other is for the test. The obtained results are ambitious.

Text Mining and Question Answering

Annotation of Cause - Result Questions in Standard Arabic Using Syntactic Grammars

Essia
Bessaies

Slim
Mesfar

Henda
Ben Ghazela

Abstract

Little research of question-answering systems has been focused on complex questions.

In this paper, we present a method for analyzing medical cause and result questions. The analysis of the question asked by the user by means of a pattern based analysis covering the syntactic as well as the morphological level.

These linguistic patterns allow us to annotate the question and the semantic features of the question and allow for extracting the focus and topic of the question.

We start with the implementation of the rules which identify and annotate the various medical named entities. Our named entity recognizer tool (NER) is able to find references to people, places and organizations, diseases, viruses, as targets to extract the correct answer from the user.

The NER is embedded in our question answering system. The task of QA is divided into three phases: question analysis, segmentation, and passage retrieval & answer extraction. Each phase plays a crucial role in overall performance.

We use the NooJ platform which represents a valuable linguistic development environment. The first evaluations show that the actual results are encouraging and could be applied to further question types.



RIADI, ENSI,
University of Manouba
[Manouba, Tunisia]



essiabessaies@gmail.com
mesfarslim@yahoo.fr
henda.benghezala
@ensi.rnu.tn



Information extraction
Medical questions
Named entities
Local grammar
Arabic language
NooJ

References

Girju, R.(2003) **Automatic Detection of Causal Relations for Question Answering**. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, Association for Computational Linguistics, pp. 76-83.

Mesfar, S. (2007) **Named Entity Recognition for Arabic Using Syntactic Grammars**. In *NLDB'07: Proceedings of the 12th international conference on Applications of Natural Language*

to *Information Systems*, pp. 305-316.



A Bottom-Up Approach for Moroccan Legal Ontology Learning from Arabic Texts

Kaoutar
Belhoucine

Mohammed
Mourchid

Samir
Mbarki

Aziz
Mouloudi

Abstract

Ontologies hold a great importance to modern knowledge-based systems. They serve as explicit, conceptual knowledge models to share a common understanding of information in a domain and make that knowledge available to information systems. However, manual construction of ontologies is an expensive and time-consuming task because of the difficulty in capturing knowledge. A solution for this issue is providing an automatic or at least semi-automatic support for ontology construction. This operation is usually referred to as Ontology Learning (OL).

Cimiano describes the tasks involved in OL as forming a layer cake. The cake is composed, in ascending order, of terms acquisition, synonyms acquisition, concepts formation, taxonomy definition, relations definition, and finally axioms definition. In order to accomplish these tasks, several ontology learning tools have been proposed in the literature. As examples: Terminae, OntoGain and Text2Onto. They differ according to input data types (format and language), output formats, and mainly the methods used in order to extract the ontological structures. Unfortunately, the Arabic language remains not supported by these tools, even it is one of the most spoken languages in the world.

In a previous work we proposed a middle-out approach for building an Arabic legal domain ontology. Accordingly, we combined two complementary strategies. The first is a top-down strategy which reuses an existing legal core ontology (LKIF-Core ontology) to define the conceptual structure of the legal domain at a language-independent level. The second is a bottom-up strategy which acquires the specific terminology from Moroccan legal texts to populate and refine the conceptual structure.



Ibn Tofail University
Kenitra
[Kenitra, Morocco]

kaoutar.belhoucine@gmail.com
mourchidm@hotmail.com
mbarkisamir@hotmail.com
mouloudi_aziz@hotmail.com



Ontology learning
Taxonomies definition
Legal field
Arabic WordNet
Arabic language
NooJ



References

Biebow, B., Szulman, S.:
TERMINAE: A linguistics-based tool for the building of a domain ontology. In Proc. EKAW '99 - Proceeding of the 11th European Workshop on Knowledge Acquisition, Modeling, and Management, Berlin, Germany, pp. 49–66, (1999).

Silberztein, M.: **Formalizing**

Natural Languages: The NooJ Approach. Wiley-ISTE, Jan. 2016, ISBN: 978-1-84821-902-, (2016).

Black, W., et al.: **Introducing the Arabic WordNet project.** In Proceedings of the third international WordNet conference, Sojka, Choi: Fellbaum & Vossen (eds), (2006).

The aim of this work is to enrich and extend our bottom-up approach in order to learn a domain-specific taxonomy from text. The new approach consists mainly of four tasks and uses the NooJ platform to implement the linguistic resources:

Corpus study: consists of lexico-syntactic analysis of Moroccan legal texts. For that, we built a domain specific dictionary (noun, adjectives) and a set of morphological grammars to perform the automatic lexical parsing of texts, then we modeled local grammars to remove lexical ambiguities.

Terms acquisition: A term can be a common noun as well as a complex nominal structure with modifiers (typically, adjectival and prepositional modifiers). To perform annotation and extraction of terms, we modeled a set of local grammars that express syntactic links of nominal groups. Furthermore, we performed a statistical step in order to identify only the relevant terms.

Concepts and taxonomy definition: most of the research in concept and taxonomy definition addresses the question from a clustering perspective, regarding concepts as clusters of related terms. For that, we elaborated a cascade of local grammars that identify the acquired terms sharing a large number of syntactic contexts. The obtained clusters guided us to define the concepts and taxonomic (hierarchical) relationships between them. At the end of this task, we have enriched the dictionary with semantic properties that reference the concepts and add their hypernyms.

Synonyms acquisition: for each concept-word of the defined taxonomy, we build a reference hypernym tree and we locate the corresponding sense within Arabic WordNet (AWN). If a concept-word has multiple senses, we get the AWN hypernym tree for each of those senses and we calculate their semantic similarity with the reference hypernym tree. The synonyms with the most similar sense to the concept have been added as a semantic property in the dictionary.

The originality of our approach lies in the semantic analysis achieved to define the concepts and the taxonomy, as well as the disambiguation phase to identify the synonyms corresponding to each concept.

Answering Arabic Complex Question

Sondes
Dardour

Héla
Fehri

Kais
Haddar



University of Sfax
[Sfax, Tunisia]



dardour.sondes
@yahoo.com
hela.fehri@yahoo.fr
kais.haddar@yahoo.fr



Complex questions
Question answering
system
Medical domain
Arabic language
NooJ

Abstract

With the increasing growth of electronic documents, the demand for Question-Answering (QA) systems is particularly great and growing, which can effectively and efficiently aid users in their information search. QA is a task that aims to automatically extract a precise answer to a natural language question.

There are many types of questions in the QA field. The most popular are the type of definition question, i.e., those asked using the word what, and the type of factoid questions, i.e., those asked using the words: how much/many, when, where, what and who.

Factoid questions usually deal with named entity, such as organization, location, person, etc., and look forward to a short identifiable answer.

Another type has to do with the questions using the words how to or why, e.g. *'Why is cancer increasing?'* or *'How to prevent hair loss?'*. These questions are less common and more challenging because they are more complex and harder to answer than factoid questions. Complex questions require a different approach than definitional and factoid questions, as their answers are not named entities, and they tend to be longer and more complex.

The aim of this paper is to propose a new approach to handle Arabic complex questions in the medical field. Our proposal is based on four components: corpus study, question analysis, documents/passages retrieval, and answer extraction.

We use dictionaries and transducers to answer any complex medical question in the Arabic language using NooJ platform. Experimentations of our Arabic medical QA system show interesting results.

References

Azmi, A., and Alshenaifi, N. (2017). **'Lemaza: An Arabic why-question answering system'**. Natural Language Engineering, 877-903.

Ezzeldin, A., and Shaheen, M. (2012). **'A survey of Arabic question answering: Challenges, tasks, approaches, tools, and future trends'**. In Proceedings of The 13th International Arab Conference on Information Technology (ACIT 2012), 1-8.

Dardour, S., Fehri, H., and Haddar, K. (2019).

**Disambiguation for Arabic
Question-Answering System.** In
International Conference on
Automatic Processing of Natural-
Language Electronic Texts with
NoJ ,101-111.

Development of a Question-Answering System in the Legal Field Based on Ontologies

Ismahane
Kourtin¹

Abdelaaziz
Mouloudi

Samir
Mbarki

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) (Kafcah Emani & Haralmbous, 2019; Abacha, 2012), 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, allowing users to ask a question on the desired information using natural language without having to browse through the documents.

An ontology of the legal field built from documents of legal decrees and laws is necessary. The purpose of this ontology is:

- (i) to formulate the interrogation requests of the Moroccan legal decrees and laws corpus, and
- (ii) to extract the response from the documents of the corpus.

In this article, we present the architecture of our Question-Answering System which consists of five main steps:

1. The creation of a corpus of legal decrees and laws: in this step we create a corpus of Moroccan legal decrees and laws in Arabic language in the form of standardized documents on which we

¹ELLIADD Laboratory,
Bourgogne-Franche-Comté University
[Besançon, France]



MISC Laboratory, Faculty
of Science, Ibn Tofail
University
[Kenitra, Morocco]

kourtin_ismahane.math
@yahoo.fr
mouloudi_aziz
@hotmail.com
mbarkisamir
@hotmail.com



*Question-answering
system – QAS
NLP*



*Legal domain
Legal decrees and laws
Legal ontology
Arabic language
NooJ*

References

Cheikh Kafcah Emani and Yannis Haralambous. "Un système de questions-réponses dans le domaine légal : le cas des

réglementations maritimes".
2019.

Asma Ben Abacha. "**Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS**". 2012.

Héla Fehri, Kais Haddar and Abdelmajid Ben Hamadou. "**A New Representation Model for the Automatic Recognition and Translation of Arabic Named Entities with NooJ**". 2011.

can apply algorithms for automatic language processing with NooJ (Fehri et al., 2011).

2. The construction of a legal ontology: next, we build a legal ontology which describes the semantic content of our corpus of Moroccan legal decrees and laws.
 3. The formulation of the interrogation requests: in this step, we analyze the question using linguistic analysis with NooJ in order to extract the question's type and patterns (subject, predicate, object).
 4. The extraction of the response document: this step of our approach consists of extracting the decree containing the answer to our question.
 5. The extraction of the answer to the question: after extracting the decree containing the answer to the user's question, we build a precise answer using our legal Ontology in response to this question.
-

AI, Lexicons and Dictionaries

NooJ for Artificial Intelligence: an Anthropic Approach

Mario
Monteleone

Abstract

Today, Artificial Intelligence (AI) is a topic mainly dealt with for commercial purposes and, from a technical-scientific point of view, often confusing the functionalities of a human-machine interface (HMI) with those a robot should have instead to imitate or replace human beings in some specific cognitive tasks. This is also caused by often inaccurate definitions of AI, which do not seem to stem from an anthropic approach, that is, which are not based on the need to accurately replicate human activities.

Actually, by definition, today AI is part of the cognitive sciences group, and it calls upon computational neurobiology (particularly neural networks), mathematical logic (as a part of mathematics and philosophy) and computer science. Similarly, it copes with stochastic-based problem-solving methods with high logical or algorithmic complexity.

On the contrary, in the current AI system, the use of Natural Language (NL) is scarcely considered, despite it being the most specific and peculiar among human qualities and characteristics. Consequently, AI today seems rather detached from the incorporation of studies on General Linguistics (GL), Rule-Based Computational Linguistics (RBCL), Morpho-syntactic formalization (MSF) (as the one produced by NooJ and Lexicon-Grammar), Formal Semantics (FS), Natural Language Processing (NLP) and Natural Language Understanding (NLU).

Therefore, the aims of our study will be to:

- 1 - Make a wide overview of the tools defined today as AI tools, focusing specifically on those that require speech recognition and interfacing with Web search engines;
- 2 - Verify the truthfulness of the information produced on AI issues by subjects with high communicative and commercial impact, such as

Dipartimento di Scienze
Politiche e della
Comunicazione,
Università degli Studi di
Salerno
[Fisciano, Italy]



mmonteleone@unisa.it

NooJ FSAs/FSTs
Artificial intelligence
Human-machine
interfaces
NLP
Lexicon-grammar
Sentiment analysis
Fuzzy logic
NooJ



References

S. Pelosi, A. Maisto, **A lexicon-based approach to sentiment analysis. The Italian module for NooJ**, in Monti J., Silberstein M., Monteleone M., di Buono M.P. (eds.) *Formalising Natural Languages with NooJ 2014*. Cambridge Scholars Publishing. International NooJ 2014 Conference, 3-5 Giugno 2014, Sassari (Italy)

M. Silberztein, **La formalisation des langues : l'approche de NooJ**. ISTE Ed., 2015, Londres (UK)

Google. In this case, our attention will focus on the issues related to linguistic automatisms, such as Automatic Translation (AT), Question-Answering (QA) and Problem-Solving (PS);

3 - Highlight the current limits of the methodological settings related to AI;

4 - Produce a state of the art scientific theories and technological innovations qualified for supporting an anthropic development of AI, such as ontologies, Fuzzy Logic, the hypothesis known as "Il Fermione di Maiorana", and quantum computers;

5 - Last but not least, highlight how the NLP tools offered by NooJ are indispensable for a correct and effective structuring of AI. To this end, we will build a set of finite-state grammars / transducers specific for Fuzzy-type problem solving, in which the answer given to a given question will also provide the "maybe" answer, together with the more classic ones of "yes" and "no". In addition, all three types of responses will be ranked using specific Sentiment Analysis (SA) settings and procedures.

New Global Cloud Solutions and NooJ's Woven Digital Intelligences for Homologated Synthetic Fixed Communication

Raffaele Marcone Rosa Giulio	Giulia Savarese Marianna Greco²	Roberto Capone Colomba La Ragione³	Javier Julian Eriquez⁴
---	---	--	---

Abstract

Smart mobility influences life of our cities with new technological solutions. An economy driven by an immense amount of data that allows us to do new business. It opens a new cultural horizon and a new education in which knowledge merges and osmotically generates new knowledge.

We explore teaching strategies and excellent skills, which can meet the needs of higher education, such as the latest generation of technologies used in business, social and personal life, as well as in institutions and colleges of higher education. Knowledge is reformulated and disciplines acquire new connotations specialized in scientific fields such as scientific and distributed informatics, and computational linguistics. The new communication belongs to the NLG languages and, therefore, the transmission is done through synthetic, standardized and non-compositional and area-based techniques. It is no longer a prescriptive, inaccurate and abstract grammar, but precise rules for the description of the codes, which, by calculation, includes categories with sentences (minimum semantic element) and then entrusts the software with the resolution in the linguistic environment.

The research questions that the team asks and describes are the following: what strategies and what reference models should the virtual communicator implement? To provide answers and validations, the team uses reference systems to arrive at scientifically valid answers, such as:

University of Salerno
[Salerno, Italy]

²MIUR
[Salerno, Italy]

³Dir Centro linguistico
Università Pegaso
[Napoly, Italy]

⁴Universitat Politècnica
de València
[València, Spain]

r.marcone7
@studenti.unisa.it

gsavarese@unisa.it
rcapone@unisa.it
rgiulio@unisa.it

colomba.laragione
@unipegaso.it

jajuen@alumni.upv.es

*Digital intelligence
Technological solutions
Teaching strategies
Skills
NooJ*

References

Adalta Wolfram Sistema

Mathematica e le tecnologie Wolfram wolfram@adalta.it
WS1.1 WS2.2; WS2. 2020

Bendjoudi, A. (2020). **Python codes for curved and regular coordinates**. Preprint.
www.academia.edu

BuViTeGMS© (2019). **System : Digital Intelligence** A.W: text editor and text reformer.

Silberztein, M. (2015). **La formalizzazione delle lingue: l'approche de NooJ**. ISTE: Londra.

- 1) Ahmida Bendjoudit's (2020) model to identify the new production parameters of fixed structures in first and second level experimental language equations.
- 2) Max Silberztein's (2015) NooJ system, for the production of analysis and paraphrases of sentences and tools to develop formal dictionaries and grammar and NLP applications such as Automatic semantics annotators, paraphrase generation.
- 3) Wolfram Technologies (2020) functionalities for information management, processing and consultation, integrated functions for advanced statistical analysis - methodologies for acquiring image information of the parts produced - with image processing and Metrics techniques for images - The implementation of advanced integrated Machine Learning algorithms for the creation of knowledge from the data obtained.
- 4) BuViTeMS (©2020) Digital Intelligence A.W. model for the production of analysis of fixed sentences and production in high-calculation environments for the production of textual paraphrases.

But the answer that intends to validate the excellence of the team must be possessed by the producer of analysis, which is the human element that produces systems and instruments. A precision team has chosen the validation systems according to the skills it possesses because the choice of models is individual.

In this paper, subsequently, we suggest a common structural design that implements new technological solutions by using precise rules for the description of the codes to obtain and present teaching strategies and excellent skills, supported by Adalta – Nooj - BuViTeMS' models as innovative tools for their users.

Standardization and Implementation of Lexicon-Grammar Tables in NooJ Platform

Asmaa
Kourtin

Mohammed
Mourchid

Abdelaaziz
Mouloudi

Samir
Mbarki

Abstract

The lexicon-grammar representation is a very important linguistic approach in automatic natural language processing. It consists of studying the lexicon of the language and all its syntactico-semantic properties. The lexicon must be classified and described by tables where the lines depict the entries, the columns depict the syntactico-semantic properties and the cells contain either a string, "+" or "-" to specify whether an expression has a property or not, or "~" if it is not yet coded (Gross, 1975). These tables are integrated into the NooJ platform in the form of dictionaries and syntactic grammars (Silberztein, 2015).

Nowadays, automation has become very important due to its advantages as optimizing time and human effort. In this regard, our contribution so far was about the automatic generation of dictionaries from lexicon-grammar tables (Kourtin et al., 2020). However, this automation suffered from the non-standardization of lexicon-grammar tables, where manual preprocessing of each table was necessary. Thus, the non-existence of a framework helping and assisting linguists to define unified properties names makes the automatic generation task difficult. Indeed, we can find two properties with the same meaning but with different names, or two properties with the same name but with different meanings. This can be an obstacle to the smooth functioning of the generation process.

In this paper, we will list the properties used in the existing lexicon-grammar tables by giving their syntactic and semantic meaning with examples, and the number of their occurrences. Then, we will propose



MISC Laboratory,
Faculty of Science,
Ibn Tofail University
[Kenitra, Morocco]

asmaa.kourtin
@yahoo.fr
mourchidm
@hotmail.com
mouloudi_aziz
@hotmail.com
mbarkisamir
@hotmail.com



Lexicon-grammar tables
Dictionary
Standardization
Syntactic grammar
Syntactic analysis
Semantic analysis
NLP
French language
NooJ



References

M. Gross. "**Méthodes en syntaxe : Régimes des constructions complétives**". Hermann, Paris, France, 1975.

M. Silberstein. "**La formalisation des langues, l'approche de NooJ**". ISTE Editions London 2015.

A. Kourtin, A. Amzali, M. Mourchid, A. Mouloudi, S. Mbarki. "**The Automatic Generation of NooJ Dictionaries from Lexicon-Grammar Tables**". In: Fehri H., Mesfar S., Silberstein M. (eds) Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications. NooJ 2019. Communications in Computer and Information Science, vol 1153. Springer, Cham (2020).

a unified meaningful name for each property to be used in the future regardless of the language.

Since the French lexicon-grammar tables are the most available at the moment, we normalized them by using the proposed unified properties. During this process, we first listed the properties of many French lexicon-grammar tables, and we obtained 510 properties. We only focused on important properties that widely appear and may be used in other languages. For example, the property "NHum" or "N=: Nhum" appears 146 times in the tables V_1, V_3, V_4, etc. It means a human or animated noun and will be replaced by "H", "H0" for the subject, or "H1" for the first complement, etc (see table below).

Prop. name before standardization	Prop. name after standardization	Meaning	Num. of occ.	Tables
NHum N=: Nhum	H, H0, H1, H2, ...	Human or animated noun	146	V_1, V_3, V_16, V_18, V_31R, V_32H, ...
N-Hum N=: N-Hum	NH0, NH1, NH2, ...	Non-human noun	96	V_1, V_3, V_5, V_16, V_18, V_32A, ...
NAbstrait NO=: Nabs	A0, A1, A2, ...	Abstract noun	10	V_4, V_31H, V_32C, V_32D, V_35LD, ...

To sum up, we tried to deal with the majority of existing properties as shown in the previous table. After giving a unified and meaningful name to each property, we will make a change to the French lexicon-grammar tables' representation in terms of their properties. This operation will facilitate more the automatic generation of dictionaries from lexicon-grammar tables since we avoid manual preprocessing of these tables. The standardization of the lexicon-grammar table will undoubtedly have a positive effect not only on the generation process but also on the efficiency of the analysis of texts and corpora.

Using Linguistic Software NooJ in Describing Preschool and Younger School Children's Croatian Language Vocabulary

Katarina
Aladrović Slovaček

Abstract

Croatian language acquisition process has been recorded in CHILDES, the world's database of children's language based on the transcription of three children's parents and so the research into Croatian language acquisition is based on the results of these very transcriptions. Some other sporadic research of kindergarten and preschool children's Croatian language acquisition has been done which has resulted in broadening the knowledge about the process of Croatian language acquisition up to the start of school education.

However, since the process does not end upon starting school but is continued parallel to the language learning process, it is important to study the course of this process and all the factors that influence it. For this reason, the project '*Developing preschool and younger school children's vocabulary*' was started with the goal to describe preschool and younger school children's vocabulary and possible influence of different linguistic and extralinguistic factors on this process.

Since most scientific data are based on estimates, for instance, that a child at the age of seven has about 10 000 words in his/her mental lexicon, it is important to precisely define the vocabulary of preschool and younger school children and establish which factors mostly affect its enrichment.

For the purpose of conducting this research, three corpora will be collected: the corpus of first to fourth grade students' written works, the corpus of selected literary works for obligatory reading for students from the first to the fourth grade and the first to fourth grade textbook corpus. A lexical and grammatical analysis of the selected corpora will be done in NooJ in order to describe a corpus on the lexical level



The Faculty for Teacher
Education
[Zagreb, Croatia]



kaladrovic@gmail.com



*Preschool children
vocabulary*
*School children
vocabulary*
Lexical diversity
Lexical density
Croatian language
NooJ

References

Aladrović Slovaček, K. i Agić, Ž. (2019) „**Umni leksikon djece mlađe školske dobi.**“ U: Um i jezik, ur. M. Matešić i A. Vlastelić, 187-200. HDPL i Filozofski fakultet Sveučilišta u Rijeci.

Kuvač Kraljević, J., Hržica, G., Olujić, M., Kologranić Belić, L., Palmović, M. i Matić, A. (2016) „**Uzorkovanje specijaliziranih korpusa govornog i pisanog jezika odraslih: izazovi i**

nedoumice.“ U: Metodologija i primjena lingvističkih istraživanja, 159-170.

Schmitt, N. 2014 **Conceptual review article – Size and Depth of Vocabulary Knowledge: What the Research Shows.** U: Language Learning Research, 64: 4, 913-951.

regarding the number of tokens, types, and both lexical density and diversity. Moreover, applying categorial analysis, tokens will be divided regarding parts of speech and the frequency of appearing in certain contexts.

Initial data obtained on these three corpora will be presented in this paper and they will serve only as an estimate of the initial state so that experimental material, aimed at enhancing vocabulary enrichment, could be prepared. The obtained data will also be compared with regard to the known extralinguistic factors: place of residence, gender and age.

Furthermore, this paper aims to describe the possibilities that using linguistic software NooJ offers in collecting, transcribing and comparing the obtained data and their lexical, grammatical and stylistic level. It is expected that the possibilities created by NooJ will facilitate to a great extent the realization of the project of describing preschool and younger school children's vocabulary.

New Russian Resources for Silberztein's Software *NooJ*

Vincent
Bénet

Abstract

Linguistic resources for the Russian language were written about five years ago, and were mostly oriented towards pedagogical uses, with grammatical annotation and tags. The feedback of users and the need for using NooJ in literary and linguistics studies suggested some improvement, which concerns above all semantics, but also linguistics and grammar.

Most researchers use the Russian National Corpus (RNC), because of the full linguistic and semantic annotation it provides. But the RNC does not allow to work with other texts, that are not part of the Corpus. The purpose of this paper is to present the improvement made in the linguistic resources for the Russian language, so that every researcher could work with NooJ with the texts they need, with full semantic and linguistics information.

The first dictionary of NooJ was based on the Zaliznyak's grammatical dictionary, which contains about 95,000 entries. But that dictionary does not have any derivational information or semantic tags. Thus, we aim to improve the NooJ Russian resources by adding all the needed **linguistic** information for derivation to NooJ dictionary, which is important in the Russian language. First of all, verbal derivation with all the prefixes and suffixes. There are a little more than 20 graphical forms of verbal prefixes, with different meanings, and each verb allows a certain number of prefixes. For Russian lexicographers, some verbs are considered to be derived verbs (with no dictionary entry) and others have a proper entry. There are some different suffixes too, and it would be very useful and very pedagogical to get an automatic tool in NooJ to transform perfective verbs into imperfective.

Russian verbs (like verbs in all Slavonic languages) have two forms, one called perfective, and one imperfective. The usual dictionary entry may vary from one lexicography to another. Some verb entries are perfective



National Institute of
Oriental Languages and
Civilizations in Paris
[Paris, France]



vincent.benet@cncrs.fr



Linguistic resources
Semantic information
Linguistic information
Russian language
NooJ

References

Tuzov, V. (2004). **Computer Semantics of Russian**, Saint Petersburg State University.

Andrey Flichenkov, Lidia Pivovarova, Jan Žižka (Eds.), **Artificial Intelligence and Natural Language: 6th Conference, AINL 2017**, StPetersburg, Russia September 20-23, 2017, Springer.

some are imperfective, and sometimes the choice may be quite surprising. In the current NooJ dictionary of verbs, we adopted the entries of Zaliznyak's: each verb has its own entry, with no relation between the verbs. As the transformation rules are rather complicated (four different graphical suffixes, vocal and consonant alternations) the verb dictionary has to be completed with the grammatical rules, which depend on the verbal category.

We also have to add all the affixes or suffixes for noun and adjectival derivation, which is widely used in Russian. The derivation indicates pejorative or diminutive (meliorative or hypocoristical) meanings of the word.

The second part of the improvement concerns **semantics**.

Some semantic tags have already been added (body parts, colors, sports, professions) for about 5,000 words in the dictionary. Some semantic criteria were added by the way of grammars, but they need improvement.

Our present work is to find a way to combine the NooJ dictionary with some information from the Tuzov's dictionary, which contains more than 150,000 lexemes with their morphological characteristics and semantic class(es). A full semantic description can be implemented in the NooJ Russian resources. Our work nowadays concerns some choices that are to be made to avoid too much ambiguities, for instance, is it necessary to add the semantic tag "food" to all vegetables? That is why we have to decide which tags must be added to the dictionary itself, and which should be added to specific grammars, that are more flexible and that any user can build.

When we successfully add derivation information for verbs (prefixes and suffixes), and semantics tags, the Russian resources for NooJ will provide a powerful tool for all kinds of research or studies.

Al-Erfan-DIC: The Electronic Dictionary of Standard Arabic Using NooJ Platform

Mohamed
El Hannach

Ahmed
Bounoua

Abstract

In this article we present the linguistic-computer techniques used in the construction of the standard Arabic dictionary *Al-Erfan-DIC*, using the NooJ platform. This platform is well-adapted to the morpho-syntactical intro-inflectional / *fusionist* system of Arabic which is different from languages based on *lemmatic* system that characterizes Indo-European languages.

It is well known that the Arabic word is formed essentially via the combination of two basic units: Root + Pattern. Roots are sequences of 3 or 4 consonants. There are about 11400 roots: 9600 trilateral roots and 1800 quadrilateral roots. Around 1036 patterns in Arabic provide the mechanism that distributes vowels on the consonants of the roots. For example, by applying the pattern “*Fa3aLa*” applied to the root “*KTB*” we obtain the word “*KaTaBa*” (the verb to write). In the same way, if we apply the pattern “*Fi3aLatun*” to the same root “*KTB*” we get the word “*KiTaBatun*” (the *maçdar* writing), etc. When applying the patterns to the roots we get more than 100 million lexical entries used and attested in this language.

In our work we bypass the traditional classification of Arabic words into verbs, nouns and particles. Here, we clearly distinguish four grammatical categories that have different syntactic, grammatical and semantic properties and are generated by separate subsets of patterns. These categories are: (V) verb, V-n (deverbal noun call *Maçdar*), V-a (adjectives) and (N) Nouns. Traditionally, the categories V-n and V-a are often neglected or confounded to N whereas they have rich grammatical, syntactic and semantic properties.

International Agency for
ANLP
[Fez, Morocco]



Faculty of Sciences,
USMBA
[Fez, Morocco]



elhannach@yahoo.com
ahmed.bounoua@gmail.com

Electronic dictionary
Root
Pattern
Database
Graph
Erfan-dic
Local grammar
Arabic language
NooJ



References

Mesfar Slim, **Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard**, Thèse, Université de Fanche Comté, France, 2008.

Misfar Slim & Silberztein Max, **Transducer minimization and information compression for Nooj dictionaries**, Proceedings of the FSMNLP 2008 Conference, Frontiers in Artificial Intelligence and applications, IOS Press, Pays-Bas, 2008.

Xantos Aris, **Apprentissage automatique de la morphologie: le cas des structures racine-schème**, Peter Lang edition, Bern, 2008.

As we can see in the following samples of application of the patterns to the roots, by just changing the infixes of a single root we can derive the four grammatical categories, for example, the root: K.T.B. we can associate with V, V-a, V-n and N.

<i>Root</i>	<i>Pattern</i>	<i>Category</i>	<i>Item</i>
<i>K.T.B</i>	Fa3ala =:	V (Verb) =:	Kataba (he wrote)
<i>K.T.B</i>	Fà3ilun =:	V-a (Adjectif) =:	Kàtibun (writer)
<i>K.T.B</i>	Fi3àlatun =:	V-n (Maçdar) =:	Kitàbatun (writing)
<i>K.T.B</i>	Fi3alun =:	N (Noun) =:	Kitàbun (book)

After completing the collection of our linguistic data (large coverage) with this approach, we started the construction of our NooJ format dictionary. We have also developed four graphs (Local grammar) for the N, V, V-n, and V-a.

Our dictionary Erfan-DIC is based on the linguistic data stored in an Access database, and put on the internet which can be consulted via the following link: www.erfan-db.com. This database (output) contains around 100 million entries described morphologically according to the NooJ notation, covering all entries of the Arabic dictionary: N, V, V-a, and V-n. All our data is stored in tables.

Given the amount of information provided by our dictionary Erfan-DIC, it is currently the basis of an application for the construction of the historical dictionary of Arabic (www.almojam.org). We will present a sample of this application in this paper.

Lexical Complexity and Basic Vocabulary of the Italian Language

Annibale
Elia

Alessandro
Maisto

Lorenza
Melillo

Serena
Pelosi

Abstract

The comprehensibility of a text depends mainly on two factors: the lexicon (chosen words) and the syntax (construction of the sentences). Also, thanks to computational tools, here we present a research that aims to assess the level of comprehensibility of the lexicon of the texts used for the linguistic education of primary school children.

Tullio De Mauro (1980) developed a vocabulary of 7050 words (Basic vocabulary), which after years of research (2016) became the New Basic Vocabulary of the Italian language (NVB) with 7,500 words. De Mauro used frequency lexicons of the written language, together with interviews to identify words that are not very frequent, but also widely used in speech. Compared to the 260,000 of the GradiT (Great Italian dictionary of use 1999 1a), De Mauro therefore selected 7500 words that can be considered basic, that is, comprehensible to a large audience, even with low levels of schooling. The NVdB is divided into words from the basic vocabulary (FO: 2000), high-use vocabulary (AU: 3000) and high availability vocabulary (AD: 2500). The NVdB marks each word with the abbreviation of the different levels.

The relationship between the lexicon of a text and the words present in the three classes of the NVdB can be used to assess the level of simplicity or complexity of the text in relation to the knowledge expected from children. For example, the high frequency in the examined text of words of the fundamental vocabulary can be considered a high index of comprehensibility, while the low frequency of the same typology of words should highlight a low index of comprehensibility.

We have computerized De Mauro's NVdB, including the brands of use, and we have included it within the Italian for NooJ form (by Simonetta Vietri). We have built a corpus made up of six complete texts of classic and contemporary fiction for children aged between 8 and 10, attending



Università degli Studi di
Salerno
[Salerno, Italy]



elia@unisa.it
amaisto@unisa.it
lmelillo@unisa.it
spelosi@unisa.it



*Electronic dictionaries
Children vocabulary
Nooj treatment
Applications to NVdB
Italian language
NooJ*

References

De Mauro T. (1980) **Guida all'uso delle parole**, n.3 dei "Libri di base", 1ª edizione, Editori Riuniti, Roma

De Mauro T. (2016) **Nuovo vocabolario di base della lingua italiana**

Lumbelli L. (2009) **La comprensione come problema**. Editori Laterza.

the last three years of primary school, consisting of approximately 312,000 tokens. The texts analyzed are: *The little Prince* by Antoine de Saint-Exupéry; *The adventures of Pinocchio* by Collodi; *Fables on the phone* by Gianni Rodari; *Jasmine in the country of liars* by Gianni Rodari; *Marcovaldo* *The seasons in the city* by Italo Calvino; *The inventor of dreams* by Ian Mc Ewan.

In our research we will present the results of the treatment with NooJ by applying the NVdB, showing the classification of the texts according to the level of lexical complexity presented.

In our opinion, this type of textual analysis can have different applications in the field of language education.

Formalising the Latin Language on the Example of Medieval Latin Wills

Linda
Mijić

Anita
Bartulović

Abstract

Language models for more than thirty languages have been created by NooJ, a tool in the field of computational linguistics. So far, such projects did not include resources for the Latin language. This paper will present an example of processing a corpus of Latin texts, the aim of which is to define a linguistic model for (medieval) Latin in the NooJ tool.

After the fall of the Western Roman Empire (476 AD), classical Latin continued to be used, and it became a *lingua franca* of medieval Europe. It was used by the Church, in science, literature, administration, diplomacy, etc. Medieval Latin underwent spelling, morphological and semantic changes, which were characteristic for all of Europe, but local variation also developed due to very specific historical and political frameworks and linguistic situations.

For the purposes of this paper, a corpus has been processed consisting of 337 wills, drawn up between 1209 and 1409 in the Zadar commune. Of these, 253 wills have been published in hard copy in *Codex diplomaticus Croatiae, Slavoniae et Dalmatiae* I–XVIII, Supplementa I–III and *Notarilia Jadertina* I–V, and 84 are manuscripts kept in the State Archives in Zadar. All the printed wills have been scanned and digitally processed in an OCR program, while the unpublished handwritten wills have been transcribed.

After a digitally readable text has been created, a dictionary has been compiled and nouns have been processed. Their semantic domains have been defined according to the function of the people mentioned in the will and the type of legacies. Based on this dictionary resource, a NooJ morphological grammar has been created, which describes the five basic types of the nominal declension system.



University of Zadar
Department of Classical
Philology
[Zadar, Croatia]



lmijic@unizd.hr
abartulo@unizd.hr



Medieval Latin wills
Morphological grammars
Latin language
NooJ

References

Silberstein, M.: **Formalizing Natural Languages: The NooJ Approach**. John Wiley & Sons (2016)

Stotz, P.: **Handbuch zur lateinischen Sprache des Mittelalters**, vol. 1–5. Verlag C. H. Beck (1996–2004)

Later work will include the development of morphological grammars to identify other types of inflected words. The resources created after the research will be made available to the wider scientific community for further research, especially in the fields of medieval studies and digital humanities.

Morphology of Middle French Verbs with NooJ

Mourad
Aouini

Laure-Anne
Caraty²

Abstract


Since 2010, the CNRS's LAMOP laboratory has been working on the PALM project, based on the name of the PALM application (*"Plateforme d'analyse linguistique médiévale"*). It was designed as a network between medievalists in order to allow the digital textual development with the textual sharing between researchers.

This application is convenient because it works with a group of metadata which allows to record a lot of information for every text. The application's library includes a textual base in Middle French with many politically oriented texts. They have been analyzed there using the NooJ approach.


One task of the PALM project is to elaborate digital linguistic tools for historians who would have little or no knowledge in Middle French, and would need tools for their work.

At the present time, we are working on the creation of inflectional digital grammar (more precisely regarding the verbal inflection). We write them with the help of NooJ, which allows us to formalize the inflection verbal pattern of certain verbs, but also to take into account the possible differences in inflectional variations. We are processing 47 verbs. They are frequent forms, extracted from the corpus of texts in Middle French, included in the library of the PALM application.

The writing of these grammars is not easy due to the state of the non-standardized French language. Moreover, at that time, it was subject to variations both diatopic, diachronic and diastratic in some cases between the 14th and the 15th centuries in France. More precisely, since the 13th century, conjugations have been gradually aligned with a single inflectional paradigm standard for each tense. Furthermore, the alternations of verbal bases, in particular for the present indicative, tend to disappear by analogy with the verbs which do not present any.

 CNRS
[Paris, France]

²University of Paris IV
[Paris, France]

 mourad.aouini@cnrs.fr
caraty.laureanne@gmail.com

Morphology
Verbal inflection
Inflectional digital grammar
Middle French
NooJ



References

Aouini, Mourad. **Approche multi-niveaux pour l'analyse des données textuelles non-standardisées: corpus de textes en moyen français**. 2018. Thèse de doctorat. Franche-comte.

Marchello-Nizia C., 2005, **La langue française aux X^{IV}e et X^{VE} siècles**. Armand Colin.

Silberstein, Max. **La formalisation des langues: l'approche NooJ**. ISTE ed., 2015.

However, if these phenomena tend to generalize, we commonly find subshape forms dating from the Old French (11th- 13th centuries ca.). In addition, the spellings are said to be “overloaded”). Several *scriptae* occur in the same text, even for verbal endings. The development of grammars must therefore take these parameters into account and leave aside certain morphological characteristics in order to develop a fairly general and operating system of reading which could be functional for all verbs in the corpus.

Our inflectional grammars currently relate only to the tenses of the indicative. We have already faced several interesting points. First, only the category of verbs in *-er* is treated with respect to simple past tense. The other groups of verbs in *-ir* sigmatic, *-ir* asigmatic, *-re* with or no alternating thematic vowel, present too much variational complexity. That is why we left these verbs aside for the moment. Moreover, if certain tenses as the imperfect or the future had rather stable and regular paradigms, this is not the case for the other ones. In other words, some tenses need new organizational groupings which do not follow the traditional classifications taught by textbooks on the history of the language.

In theory, we separate verbs into two categories at the present tense. But this tense requires ten digital grammars according to the types of ideograms and person's morphograms required for certain verbs and not for others. All these observations lead to different types of classifications between pure linguistic and computer linguistic or formal linguistic, when you start working in the field of digital humanities. This makes the use of NooJ particularly effective in the specific case of a language which does not yet obey the normative grammar, but only the rule of oral and written uses. What's more, these uses change from one territory to another.

A Morphological Grammar for Modern Greek: State of Art, Evaluation and Upgrade

Lena
Papadopoulou

Elina
Chadjipapa²

Abstract

Modern Greek NooJ Module was born in 2007 (Gavriilidou et al. 2008). Since then, a series of lexicographical data as well as morphological and syntactic grammars (Gavriilidou et al. 2010, 2013, 2015) have been imported in order to improve the results of Greek language automatic processing. Meanwhile language learning is considered one of our main purposes.

As part of this project, Greek NooJ linguistic information has been tested with a goal to validate the data. For that purpose, a NooJ corpus has been compiled. The entire text databank of the “Centre for the Greek Language” formed the source of the corpus, which includes 6 text files in total, one for each language level, corresponding to 333 text units and 128.694 tokens.

Our validation test focused on simple word units, and especially nouns. First, annotations referring to nouns were tagged and exported through linguistic analysis. The Greek NooJ dictionary and inflectional grammar were applied as resources. Subsequently, 3464 lexemes annotated as nouns along with their corresponding inflectional properties codification form the base on which the manual validation process relied.

The process that has been followed was six fold:

- (a) inflectional paradigms were revised; or rather, they were simplified as far as their codification and their internal structure,
- (b) the correspondence between inflectional codification and inflectional paradigms was checked,



Hellenic Open University
[Patra, Greece]

²Democritus university of
Thrace
[Komotini, Greece]



papadopoulou.lena
@gmail.com
elinaxp@hotmail.com



Dictionary
Morphology
Inflection
Language learning
Nouns
Greek language
NooJ

References

Gavriilidou, Z., Papadopoulou, E., & Chatzipapa, E. (2013). **Numeral-noun and numeral-adjective construction in Greek**. En M. Silberztein, A. Donabédian, & V. Khurshudian, *Formalising Natural Languages with NooJ*. Cambridge Scholars Publishing, pp. 113-122.

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

Gavriilidou, Z., Papadopoulou, E., Chadjipapa, E. 2010. **New data in the Greek NooJ module: A local grammar of proper nouns.** In Silberstein, M.; Varadi, T. (eds): Proceedings of the 2008 International Conference (Budapest). Cambridge Scholars Publishing, pp. 93-100

(c) ambiguous forms regarding the parts of speech were disambiguated,

(d) nominal word forms that are exclusively located within multiword units were deleted and they are exclusively considered as part of the corresponding phrasemes,

(e) non-active paradigms were deleted, and, all those steps at the same time as

(f) working towards the alignment with inflectional paradigms categorization proposed by Dictionary of Standard Modern Greek .

Through that process our main aim has been achieved; the aim of a dynamic morphological grammar for simple nouns based both on primary and secondary lexicographic resources. Such dynamics are referred both to the Greek inflectional grammar as well as to the Greek NooJ dictionary. The introduction of new lemmas and inflectional paradigms is performed in a systematic and unambiguous way, as it was carried out in case of the 2661 unknown tokens that were found through the linguistic analysis of our corpus.

Syntax and Semantics

Detecting Negation Scope with NooJ

Gaurish
Thakkar

Jeremy
Barnes²

Nives
Mikelić
Preradović

Abstract

The objective of this paper is to describe the methodology for explicit negation detection from the reviews taken from SFU corpus (Konstantinova et al 2012) using NooJ (Silberztein 2003). NooJ is a linguistic development environment and corpus processor which has been used to perform various levels of Natural Language Processing tasks.

Finding negation cues along with their scope is a crucial subtask in sentiment analysis (Barnes 2019) and information extraction (Morante et al 2009). According to a recent study (Koza 2018), the negated finding in medical reports can be improved by using lexical and syntactic rules using the formal grammars created in NooJ. Previous studies (Tanushi et al 2018) have reported techniques that used features like cue phrases (NegEx (Chapman 2001b), PyConTextNLP (Chapman 2011) radius of a negation cue and dependency features (SynNeg)). The scope of negation is usually captured using a window of 5 tokens.

In our study, we use the NooJ grammar to capture the negation cues and subsequently capture the negation scope if they exist in the given corpus.

For example, in the noun phrase *Magical boyhood adventures, (without) {all the cotton-picking drudgery of Grisham}* the negation cue is represented with the preposition "**without**", while the phrase "**all the cotton-picking drudgery of Grisham**" represents the negation scope.

The methodology is as follows:

1. The given text is normalized, usually involving the expansion of contractions like "can't".

Faculty of Humanities
and Social Sciences
[Zagreb, Croatia]



²Language Technology
Group, University of
[Oslo, Norway]



gthakkar@m.ffzg.hr
jeremycb@ifi.uio.no
nmikelic@ffzg.hr



Negation
Grammar
Syntactic rules
English language
NooJ

References

Koza W. et al. (2018) **Automatic Detection of Negated Findings with NooJ: First Results**. In: Silberztein M., Atigui F., Kornysheva E., Métais E., Meziane F. (eds) *Natural Language Processing and Information Systems*. NLDB 2018. Lecture Notes in Computer Science, vol 10859. Springer, Cham

Tanushi, Hideyuki, et al.
"Negation scope delimitation in clinical text using three

approaches: NegEx, PyConTextNLP and SynNeg."
19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22-24, 2013, Oslo, Norway. Linköping University Electronic Press, 2013.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001b). **A simple algorithm for identifying negated findings and diseases in discharge summaries.** JBiomed Inform, pages 34–301.

2. Regular expressions are used to find negation cues present in the given sentence. The negation is a precompiled list of words/phrases.
3. Regular expressions are used to capture negation cues.
4. The performance of the rules is tested with other algorithms like Negex.

The metric of Precision-Recall and F-1 score is reported for each of the techniques and datasets. Finally, the negation cue detection and scope detection are presented as Grammar in NooJ for negation detection.

The Annotation of Indonesian Reduplications with NooJ

Prihantoro

Abstract

This paper is a part of my on-going PhD project in the department of Linguistics, Lancaster University. Retrieving Indonesian reduplications automatically is a complex task due to 1) **reduplication-affixation interface**, including their morphophonemic, 2) **variety of reduplication forms** (full, partial, imitative) and functions (e.g plural, manner, iterative, reciprocal etc.).

So far, reduplication is not part of the annotation scheme of the majority of existing Indonesian POS taggers (Alfan & Purwarianti 2009; Rashel et al 2014). The state of the art morphological analyser of Indonesian, Morphind (Larasati et al. 2011) tokenises all reduplications like non-reduplication forms, which causes retrieval from its annotation output very challenging. Users have to rely on the plural tag, which under Morphind's scheme, is encoded to all reduplications regardless of the actual functions.

The starting point of this paper is a morphologically annotated corpus of Indonesian, whose annotations are currently implemented in NooJ by using Indonesian morphology module. The scheme of the implementation is defined in Prihantoro (2019). The module that I am building largely handles morpheme analysis on single word level. Words are tokenised into morphemes and each morpheme is linked to their formal (e.g. prefix, root, and clitics) and functional analytic tags (voices, outcome POS, aspects, etc). I am now experimenting on how to handle reduplications.

The unproductive reduplications, Imitative and partial reduplications are hard-encoded in the dictionary entry; this includes imitative and partial reduplications. Their entries are given special tag +UNAMB to prevent root entry analysis from being applied. The most productive reduplication, full reduplication, is handled differently, by using a syntactic grammar. In Indonesian, the elements of full reduplications are

Lancaster University
[Lancaster, England]



Universitas Diponegoro
[Semarang, Indonesia]



prihantoro2001
@yahoo.com

*Reduplication
Dictionary
Morphology
Disambiguation
Indonesian language
NooJ*



References

Larasati, S.-D., Kuboň, V., & Zeman, D. (2011). **Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus.** *Systems and Frameworks for Computational Morphology* (pp. 119-129). Zurich, Switzerland: Springer Berlin Heidelberg.

Prihantoro. (2019). **A new tagset for morphological analysis of Indonesian.** *International corpus linguistics conference.* Cardiff.

Silberztein, M. (2016). **Formalizing Natural Languages**

Nooj Approach. London: John Wiley and Sons.

usually separated by a hyphen as in *pelan-pelan* ‘slowly’ from *pelan* ‘slow’. The grammars are written with identity constraint to verify whether the two roots separated by ‘-’ are identical. The grammars replace the root tag in the second segment with reduplication tags.

The trickiest part is handling reduplication-circumfix interface such as *berpukul-pukulan* ‘to hit one and each other’. On the one hand, the circumfix *ber-an*, in this case, is a reciprocal circumfix attached to reduplicated root *pukul* ‘to hit’.

As NooJ .nom grammar cannot (1) accept a hyphen, and (2) perform equality check, I offer a solution which contains two steps. Step 1: ambiguously annotate the affixes as circumfix and prefix+suffix combination by using .nom grammar. This significantly increases the number of ambiguous analyses, not only for the intersection, but also with words produced by pure affixation. Step 2: perform disambiguation by using several .nog grammars. It takes 7 .nog grammars to satisfactorily perform disambiguation on all expected ambiguities, including the removal of *orphan* annotations. This solution will be tested on a larger corpus. Other solutions are also considered.

Another more straightforward solution is to encode the full morphological analyses of all reduplication-circumfix intersections in the dictionary. However, this will be an exhaustive process, as such reduplication is productive. Another solution requires the modification of NooJ Engine to accept a hyphen in .nom grammar. This will significantly reduce the complexity of .nog grammars into 2 grammars instead of 7 as in the current solution.

The Automatic Paraphrasing of Arabic Syntactic Structures

Ali
Boulaalam

Azeddine
Rhazi²

Abstract

The general objective of this paper is to exploit the existing Lexicon-Grammar Tables, as well as assessing their relative importance vis-à-vis the concept of transformation. As a matter of fact, the modifications applied to a sentence when binary and unary structural transformations performed are local (Silberstein, 2018). These operations include multiple processes at the morpho-syntactic and semantic levels, the potential swapping elementary support verbs (Mota, 2018) as well as symmetrical restructuring.

Our proposal is to model highly productive phenomena of the Arabic language such as

- nominalization (verb subjects),
- passivation and
- negation,

by performing formal operators (+) and (-), in order to formalize the relation between structures and their semantic properties and thus to represent the symmetry between sentences that share a predicate that links the noun and a support verb, which can also be an adjective or an auxiliary verb.

Furthermore, the automatic process of paraphrasing involves both the distributional and transformative features of each class of verbs or other structures, taking into account their degree of distributional semantic homogeneity.

This research outlines how to build Lexicon-Grammar Tables by using automatic paraphrasing in a large transformational grammar of the Arabic language (Elhannach 1988), on the one hand, and how to

Mly Ismail University
[Meknes, Morocco]



²Cadi Ayyad University
[Marrakech, Morocco]



lingdroit@gmail.com
ourzagh@gmail.com

*Arabic syntactic analysis
Transformations
Lexicon-grammar tables
Automatic paraphrasing
Arabic language
NooJ*



References

Mota Cristina, Baptista Jorge and Barreioro A, **The Lexicon-Grammar of Predicate Nouns with ser de in Port4NooJ**, Formalizing Natural Languages with NooJ 2018, and Its Natural Languages Processing Applications, Communications in Computer and Information Science, 987.Springer 2018.

Sagot, B., Tolone, E.: **Intégrer les tables du lexique-grammaire à un analyseur syntaxique robuste à grande échelle**. In: Actes de la conférence TALN 2009.

Silberstein, M. Unary
**Transformations for French
Transitive Sentences,
Formalizing Natural Languages
with NooJ** 2018, and Its Natural
Languages Processing
Applications, Communications in
Computer and Information
Science, 987. Springer 2018.

implement Lexicon Grammar Tables to expand both NooJ electronic
dictionaries and grammar rules on the other hand.

Using NooJ for Marketing Choices

Magali
Bigey

Abstract

The aim of this proposal is to show and explain a system to analyze marketing choices for watches brands, with NooJ and its tools, as one of a few other means for representation of the audience of the brand CODE41.

CODE 41 is a Swiss watch brand founded in 2016 with a crowdfunding action. Their team wants to be accessible and transparent, playing with reception and communication based on aggressive marketing: dozens of emails, community effects, pre-order procedures and limited editions. In this presentation, we will focus on DAY41 project, by a semio-linguistic analysis and an audience reception on marketing.

For almost three years, CODE41 has broken the codes of Swiss watchmaking by looking to be affordable while producing models of Swiss quality, with transparent manufacturing and creation but at a lower price compared to luxury watchmaking. The idea of CODE41 is then to democratize the luxury watch, doing so with direct marketing, even going so far as consulting its public and its potential customers about the creation of the models to come. We will see how it is working, creation of a lead generation and a “CODE41 community”, built around very targeted and effective e-marketing, allowing it to do without distributors and sell watch models worth several thousand euros remotely. A very interesting fact is that the customers did not see it before.

With NooJ we will analyze how the brand makes itself known and how it communicates about itself, in an eight months corpora of emailing and communication on social networks. Studying consumers’ ethos makes it possible to see how they feel about the brand, without saying it. NooJ is the key to analyze each comment and the feeling of consumer, making it easier to direct the communication choices.

 ELLIADD
Université de Franche-Comté
[Besançon, France]

 magali.bigey@gmail.com

 *Brand audience*
Marketing choices
Ethos analysis
Semio-linguistic analysis
NooJ

References

Amossy R. (2010), **La présentation de soi**. Ethos et identité verbale, Presses Universitaires de France, 235 pages.

Bigey M. (2019), **La publicité horlogère 4.0 : sémiolinguistique et réception d’une image en mutation**, in « **Publicité horlogère 4.0 : les nouveaux codes** », dir. F. Courvoisier & K. Zorik, 22èmes Journées internationales du marketing horloger, Editions Loisirs et pédagogie, Le Mont-sur-Lausanne, pp. 55-62.

Hagel III J. & Armstrong A.
(1996), **The Real Value of On-Line Communities**, in Harvard Business Review (May-June), pp.134-141.

Thanks to NooJ and concordances and frequencies analysis, we can locate which are the levers or brakes for generating sales, brand awareness, circulation of shared values... NooJ can here become an interesting marketing way for linguists interested in brands, and for marketing specialists.

Computational Modeling of a Nominal Ellipsis Grammar for Spanish

Walter
Koza

Hazel
Barahona

Abstract

Ellipsis is a method to avoid lexical redundancy. This phenomenon presents a challenge both for theoretical and computational studies, since it sets up a breaking point between tracking the form/meaning in the grammatical description and automatic detection. Ellipsis has been studied in Generative Grammar from different theoretical perspectives that aim to account for the generative methods (Merchant, 2019). In the field of computational linguistics, the studies focus on data recovery from a corpus (Hardt, 1997).

However, there is little discussion between these two perspectives. Consequently, this work presents a descriptive/explanatory proposal, which accounts for the generative methods of structures with nominal ellipsis (NE), towards a computational modeling. To this end, the contexts in which elided items are perceived are analyzed and two predominant structures are established:

1. Ellipsis of the noun phrase (NP) head, object of a compound determinant phrase (DP): *El hijo de John y el de Peter* / the son of John and the son of Peter 'John's son and Peter's son'.
2. Ellipsis of a NP head, object of a DP, argument of a predicate: *El presidente de Perú homenajó al de Francia* / The president of Peru honored the president of France 'The Peruvian President honored to the French president'.

From then on, the following formal descriptions for these cases were established:

1. $[[\text{PRENOM}]1 \text{ N } [\text{POSTNOM}]1]\text{SDI Coord } [[\text{PRENOM}]2 \text{ N } [\text{POSTNOM}]2]\text{SDII}$



Pontificia Universidad
Católica de Valparaíso
Proyecto FONDECYT
1171033
[Valparaíso, Chile]



walter.koza@pucv.cl
hhgg09@yahoo.es



Nominal ellipsis
Computational modeling
Generative grammar
Spanish language
Noo!

References

Merchant, J. (2019). **Ellipsis: A survey of analytical approaches.** Ms., University of Chicago. For Jeroen van Craenenbroeck and Tanja Temmerman (eds.), Handbook of ellipsis, Oxford University Press: Oxford.

Hardt, D. (1997). **An Empirical Approach to VP Ellipsis.** Comp Linguist. 23.

2. [DET[[PRENOM]1 N [POSTNOM]1]]_arg0 PRED
[DET[[PRENOM]1 N [POSTNOM]1]]_arg1

Taking formal description as basis, the following retrieval rules are proposed.

- Given a coordination A^B , where B is a DP of an equivalent structure to A, and B presents an ellipsis of its NP head:
 1. Copy the root of the head and gender of the NP, object of A
 2. Copy number of B's Det
- Given a predicative argument structure A: $P(\alpha, \beta, \gamma, \text{and so on.})$ with an argument to the right of α that presents an ellipsis of its NP head:
 1. Copy the root of the NP object of α
 2. Copy the number of β or γ 's determiner, and so on (according to the elided argument)

Both the description and the rules were modeled with Nooj. To this end, an electronic dictionary with the entries of *Diccionario de la Lengua Española* (DLE), to which morphosyntactic data was added, was created. Then, computer grammars for NE recognition and retrieval of the elided item were developed. Context-dependent grammars were developed, these grammars set up a group of variables that can be grammar categories (N, A, V); terminal elements, such as vocabulary items or an embedded grammar. The methodology was tested in a corpus with 200 sentences and the following results were obtained:

- Coverage: 82%
- Precision: 100%
- F Measure: 90, 10%.

Errors and omissions found were due to units not listed in the electronic dictionary, as well as the complexity of certain structures.

Teaching with NooJ

Automatic Treatment of Causal, Consecutive and Counterargumentative Discourse Connectors in Spanish: a Pedagogical Application of NooJ

Andrea
Rodrigo

Silvia
Reyes

María Andrea
Fernández Gallino²

Abstract

Our intention, within the framework of the pedagogical application of NooJ to Spanish teaching undertaken by the research team Argentina, is to continue with the application of discourse tags, as we have been doing in previous work dealing with causal connectors in Spanish (Rodrigo et al 2019). We added counterargumentative and consecutive discourse connectors to our dictionary taking into consideration a specific population of Spanish learners, whose mother tongue is Spanish, but who sometimes face the same difficulties as the ones experienced by learners of Spanish as a foreign language (SFL). The category 'connector' is defined following Martín Zorraquino and Portolés (1999), who state that a connector is “a discourse marker that relates semantically and pragmatically a discourse member with another discourse member”. The corpus comprises stories for kids written by students of two Primary Education Tertiary Schools for Teachers (Escuela Normal Superior N°35 “J. M. Gutiérrez” and Escuela Normal Superior N°36, “M. Moreno”) during a Workshop on Texts Comprehension and Production. Those students signed their consent so that their children’s tales could be analysed by our research team using the NooJ platform. Some of our conclusions will be shared in this paper.

Generally speaking, the corpus texts show enormous deficits. One of the main inconveniences refers to the lack of resources to enable



Universidad Nacional de Rosario
[Rosario, Argentina]

²Universidad Tecnológica Nacional
[Rosario, Argentina]



andreafrodrigo@yahoo.com.ar
sisureyes@gmail.com
mandrea.fernandezg@gmail.com



NLP
Pedagogy
Causal connectors
Consecutive connectors
Counterargumentative connectors
Language teaching
Spanish language
NooJ

References

Rodrigo, A., Reyes, S., Mota, C., Barreiro, A. **Causal Discourse Connectors in the Teaching of Spanish as a Foreign Language (SLF) for Portuguese Learners Using NooJ**. In: Fehri H., Mesfar S., Silberztein M. (eds) *Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications*. NooJ 2019. *Communications in Computer and Information Science*, vol 1153, pp. 161-172. Springer, Cham (2020).

Rodrigo, A. and Bonino, R.: **Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua**. Ed. Ciudad Gótica, Rosario (2019)

Silberztein, M.: **Formalizing natural languages: The NooJ approach**. ISTE Wiley, London (2016)

Zorraquino, M.A. and Portolés, Lázaro, J.: **Los marcadores del discurso**. In: Bosque, I., Demonte V. (eds.) *Gramática descriptiva de la lengua española*, pp. 4051-4213. Tomo III. Espasa Calpe, Madrid (1999).

metalinguistic reflection, which in turn, in our teaching experience, is the motor of linguistic knowledge. Another difficulty tangentially lies in the lack of clarity in terms of the correction and self-correction of texts, since there is a belief that writing is the result of inspiration and that it is done all at once. NooJ can give at least a partial answer to these issues, because dictionaries and grammars contribute to the formalisation of linguistic knowledge. For this, we follow Silberztein (2016).

In our work plan, we first introduced new tags in our dictionary of discourse connectors, [C+consec] and [C+contrarg], in order to analyse consecutive and counterargumentative connectors respectively. Corpus analysis showed that causal connectors [C+caus] greatly outnumber consecutive and counterargumentative connectors. In this study, and as a result of the corpus analysis, it is deduced that, after proving their scarceness, it is necessary to strengthen the use of consecutive and counterargumentative connectors through exercises aimed at overcoming difficulties when writing in Spanish.

Finally, our idea is to show some syntactic grammars created with the NooJ platform in order to analyse corpus phrases and make visible the use of causal, consecutive or counterargumentative connectors. Our intention is to enable reflective thinking about those structures, as for example:

- *a Pablo le encantaba ir al parque pero había un problema*
(Paul loved going to the park but there was a problem)
- *pues yo si tengo amigas pero también prefiero estar sola*
(well I do have friends but I also prefer to be alone)

In all these cases, we are dealing with phrases lacking punctuation or capitalisation, commonly seen in text speak, and having orthographic flexibility as well. Connectors then acquire almost an extraordinary role in making what is expressed fairly intelligible. How can we teachers contribute to improve these texts? We believe that metalinguistic reflection can be a key in this regard. And at this point, NooJ constitutes an ideal tool. We believe that it is possible to encourage metalinguistic reflection through NooJ by proposing exercises that can be implemented in language teaching (Rodrigo and Bonino 2019).

Preparing the NooJ German Module for the Analysis of a Learner Spoken Corpus

Mirela
Landsman Vinković

Kristina
Kocijan

Abstract

This project intends to merge the knowledge of a foreign language teaching specialist and a language processing specialist to detect and classify different types of errors found in the spoken classroom discourse of Croatian learners of German as a foreign language. The authors intend to design a model within the NooJ as an NLP tool that can successfully be used for the analysis of a learner corpus (Granger, 2004). The role of NooJ in this project is to detect and annotate different types of errors in order to facilitate the analysis of such data.

The preliminary corpus consists of 5 different class interactions of Croatian master students of teaching German as a foreign language that have been recorded during their regular classes held at the Faculty of Humanities and Social Sciences at Zagreb University.

The length of each microteaching session is approximately 6 - 12 minutes while the topics covered in the classroom are diverse and were predefined by the school curriculum. The recordings have been faithfully transcribed, keeping the mistakes originally introduced by the student-speaker. Except for the student-speaker gender, mother tongue and years of language learning, no other personal data about the speaker is kept.

In order to evaluate the student-speaker linguistic, lexicosemantic and pragmatic competences, the rated criteria fall into the general linguistic range, vocabulary range, vocabulary control, grammatical accuracy, coherence and pragmatic appropriateness. This classification conforms to the recommendations by the Common European Framework of Reference for Languages (CEFR).

The project will use the main NooJ resources prepared for German (Müller, 2014) and will enlarge the dictionary with the unknown words



Faculty of Humanities
and Social Sciences
[Zagreb, Croatia]



mlandsma@ffzg.hr
krkocijan@ffzg.hr



Student learner corpus
Detecting errors
Language teaching
German language
NooJ

References

Sylviane Granger. **Computer Learner Corpus Research: Current Status and Future Prospects.** In Applied Corpus Linguistics, 123-145, 2004.

Müller R. 2014. **NooJ as a Concordancer in Computer-Assisted Textual Analysis. The case of the German module.** Formalising Natural Languages with NooJ 2013: Selected Papers from the NooJ 2013 International Conference. (eds.) S. Koeva, S. Mesfar and M. Silberstein. Cambridge Scholars

Publishing, Newcastle., UK: 197-
208

from the preliminary learner corpus prepared by the authors. Morphological grammar will be constructed to recognize any misspellings, while syntactic grammars will be introduced to find and annotate the errors on the higher levels of syntactic analysis.

Different Shades of Verbs

Formalizing Ukrainian Verbs with NooJ

Olena
Saint-Joanis



INALCO
[Paris, France]



alena.saintjoanis
@gmail.com



Aspectual pairs
Verbs
Ukrainian language
NooJ

Abstract

The Ukrainian verb, from the point of view of its morphologic structure is made up of a root, with an ending attached to it, but also possibly the suffix, prefix, and postfix – “s’a”. It is important to note that the prefix, the root, and the suffix form a stem of the verb. Each Ukrainian verb has two stems, the infinitive stem and the present stem (or non-past stem), which can happen to be the same for some verbs or slightly different for others, due to the alternation of consonants and vowels or a varying suffix.

Ukrainian verbs have syntactic characteristics, because they play the role of a predicate or a subject and show transition (transitive and intransitive verbs). We can also classify them according to semantic characteristics (motion, position, state, beginning, specific development or termination of action and others). And finally they form the morphological categories described in this table:

Morphological category	Form	Expressed by
Person	1, 2, 3	Ending
Number	Plural, Singular	Ending
Gender	Masculine, Feminine, Neuter	Ending
Voice	Passive, Active	Postfix “-s’a” or syntactic structure
Mood	Indicative, Imperative, Infinitive	Ending
	Conditional	suffix + ending + particle “by”
	Gerund	Suffix
	Participle	suffix + ending
Tens	Present	present stem of imperfective verb

References

Bogacki K., Gwiazdecka E. (2012). **Derivational Structure Of Polish Verbs And The Expansion Of The Dictionary.** In Formalising Natural Languages With NooJ: Selected Papers From The NooJ 2011 International Conference (Dubrovnik, Croatia).

Gorpynyč, V.O. (2004). **Morphologiya Ukraïnskoï Movy.** [Morphology Of The Ukrainian Language]. Akademiya, Kyïv.

Hetseвич Y. And Hetseвич S. (2012). **Overview Of Belarusian And Russian Electronic Dictionaries And Their Adaptation For NooJ.** In Formalising Natural Languages With NooJ : Selected Papers

From The Nooj 2011
International Conference
(Dubrovnik, Croatia).

Silberstein, Max (2015). La
Formalisation Des Langues.
L'approche De Nooj. Université
Of Franche-Comté, Iste Edition.

Šojat K., Kocijan K., Filko M.
(2018). **Processing Croatian
Aspectual Derivatives.** In
Formalising Natural Languages
With Nooj : Selected Papers
From The Nooj 2018 (Palermo,
Italy).

		+ ending
	Past	stem of infinitive + suffix + ending
	Future	<ul style="list-style-type: none"> ▪ verb imperfective in infinitive mood + ending ▪ analytic (2 words : present stem of verb "buty" + ending, together with infinitive of the imperfective verb ▪ stem of perfective present + ending
Aspect	Aspectual pair (Imperfective/Perfective)	suffix, prefix, place of stress

We decided to introduce the non-pair verbs in the Nooj dictionary, and then those verbs that form the pair – only imperfective verbs, and also display the aspectual pair for each verb in the dictionary. Then we describe paradigms (77 for imperfective verbs and 7 for perfective non-pair verbs), derivation for verbs accepting postfix –“s’ia”, derivations for perfective verbs form the aspectual pair (actually 27) and connect our verbs in the dictionary to these paradigms and derivations.

We achieved a fairly satisfying result, because now when we search for the occurrence in the corpus, we can link two aspects of the verb. Therefore, we would like to show the steps of our work, talk about the challenges encountered and submit the results. It should be noted that this work is a part of the production process of the Ukrainian module for Nooj that will benefit as a pedagogic tool for Ukrainian studies.

Transformations and Paraphrases for QU Sentiment Predicates

Maximiliano
Duran

Abstract

In this paper, I present a study on the possibilities of the automatic generation of paraphrases of Quechua sentiment predicates.

I show how the transformational NooJ engine could be used to produce all paraphrases of any Quechua direct transitive sentence.

First, I construct some grammars of elementary transformations, then those of Quechua direct predicates and, more specifically, of sentiment predicates.

I describe in detail some of the paraphrasing grammars: pronominalize, reduction and permutation of the arguments, and the passivization. I show how each of these can be combined with each other following certain syntactic constraints.

Finally, I show how I apply this system in order to get the automatic translation of a French direct transitive phrase into Quechua.



Université de Franche-Comté
[Besançon, France]



duran_maximiliano
@yahoo.fr



Syntactic analysis
Transformational analysis
Transformational grammar
Sentiment verbs
Sentiment predicates
Paraphrase
Automatic translation
Quechua language
NooJ

References

Duran, M., **The annotation of compound Suffixation Structure of Quechua Verbs**. In: Proceedings of the 2014 International Conference and Workshop, Sassari (2014). Springer, Switzerland (2016)

Duran, M., **Morphological and syntactic grammars for recognition of verbal lemmas in Quechua**. In: Proceedings of the

2014 International Conference
and Workshop, Sassari (2014).
Cam-bridge Scholars Publishing,
Newcastle (2015)

Silberztein, M.: **Automatic
transformational analysis and
generation**. In: Gavriilidou, Z.,
Chatzipapa, E., Papadopoulou, L.,
Silberztein, M. (eds.) Proceedings
of the NooJ 2010 International
Conference and Workshop, pp.
221-231. Univ. of Thrace,
Komotini (2011)

Verbal syntax in Rromani: Diatheses

Masako
Watanabe

Abstract

This paper aims at presenting a morphosyntactic study concerning the diatheses (i.e. the active, the passive and the reflexive passive) in Rromani in order to program the verbal syntax of the Rromani language in the NooJ system.

The basic word order in Rromani is SVO, however, the person of the subject is marked on the verbal ending and the personal pronoun of the subject is often deleted except if it is necessary to specify or insist who it is about. **Khoses i mez.** (*You wipe the table.* **Tu khoses i mez.** *You (that is not other persons) wipe the table.* / **YOU** wipe the table.

At the morphological level, there are two diatheses in Rromani each of which is expressed by inflectional forms: the active and the medio-passive. The active indicates the action that the object undergoes, while the medio-passive is the action that the subject undergoes, either by his own means or by other persons. This means that there can be a semantic ambiguity with medio-passive forms (i.e. the reflective or non-reflective passive). **Khoslös.** (*You wipe yourself.* / *You are wiped (by someone else).* The form **khoslös** has been programmed in NooJ inflectional morphology.

However, there are analytical forms equivalent to each of the semantic values: the reflective passive is expressed by the active and the person of the subject in oblique form (i.e. the object complement), while the non-reflective passive is expressed by the copula and the verb in the past participle passive. **Khoses tut.** (*You wipe yourself.* **Sinan khoslo.** (*You 'male' are wiped (by someone else).*

Then, the personal pronoun in the third person (singular and plural) in the oblique and reflexive form (**pes** *himself/herself*, **pen** *themselves*) is distinct from that of a non-reflexive (**les** *him/her*, **len** *them*) in Rromani. So there is no risk of confusing the persons of the object. **Khosel pes.**



University of Franche-Comté
[Besançon, France]



masakowatabe@free.fr



Diatheses
Medio-passive
Reflexivity
Transitivity
Causative
Rromani language
NooJ

References

Courthiade, M. et al.: **Morri angluni rromane čhibăqi evroputni lavustik.** Romano Kher, Budapest (2009).

Courthiade, M.: **Xaca dume - but godi.** MS.

Silberstein, M.: **La formalisation des langues: l'approche de NooJ.** ISTE Eds., London (2015).

(He/She) wipes himself/herself. **Khosel les.** (He/She) wipes someone else.

Regarding the transitivity in Rromani, there are two types: the transitive and the intransitive. **Phabares o kašt.** (You) burn the wood. **O kašt phablöl.** The wood burns. Verbs in Rromani cannot be both transitive and intransitive as is the case with the verb to burn in English for example.

The causative can be formed with a suffix (e.g. **-är-**, **-av-**), for example, **khosel** to wipe > **khoslärel** to make wipe, **siklöl** to learn > **siklärel** to teach (i.e. to make learn), **daral** to be afraid > **daravel** to scare. Now, **siklöl** to learn is a medio-passive form of the verb **sikavel** to show which already has a suffix of the causative **-av-**. On the other hand, the form without the suffix ***sikel** does not exist.

In NooJ dictionary for Rromani, the inflectional forms (e.g. **khoslöl** to wipe himself / to be wiped), nor that of derivation (e.g. **khoslärel** to make wipe) are not generally found as entry words, on the other hand, if they are lexicalized (e.g. **siklöl** to learn, **siklärel** to teach), they are treated as entry words.

Through a morphosyntactic study, we will build templates of verb phrases especially the diatheses in order to develop NooJ syntax for the Rromani.

Arabic Psychological Verbs Recognition through NooJ Transformational Grammars

Asmaa
Amzali

Mohammed
Mourchid

Abdelaziz
Mouloudi

Samir
Mbarki

Abstract

The Psychological verbs are an important part of a language vocabulary. They are widely used in newspapers texts, social networks messages (Facebook, Twitter), novels, etc. This makes syntactic-semantic analysis of texts containing psychological verbs very useful in NLP applications (sentiment analysis, question answering systems, etc.). In fact, such analysis makes possible to know people's concerns, their feelings, their tendencies, etc. Thus, it can help those in charge with their decision-making.

In order to survey the opinion of Moroccan youth regarding their daily interests and concerns, we have created a corpus of journalistic texts extracted from different Arabic newspapers.

Based on the dictionary and grammars developed in Amzali et al. (2020)], in this paper we have carried out an analyzer which recognizes different psychological verbs in this corpus, even if a lot of them do not appear only in their basic form. In fact, they can appear in transformed sentences like passive, negative, nominative, etc., such as in the sentence

‘يُخَافُ مِنْ قُوَّتِكَ’ (yokhaafo min kowatika - fear from your power)’.



MISC Laboratory, Faculty
of Science, Ibn Tofail
University
[Kenitra , Morocco]

92



asmamzali
@hotmail.fr
mourchidm
@hotmail.com
mouloudi_aziz
@hotmail.com
mbarkisamir
@hotmail.com

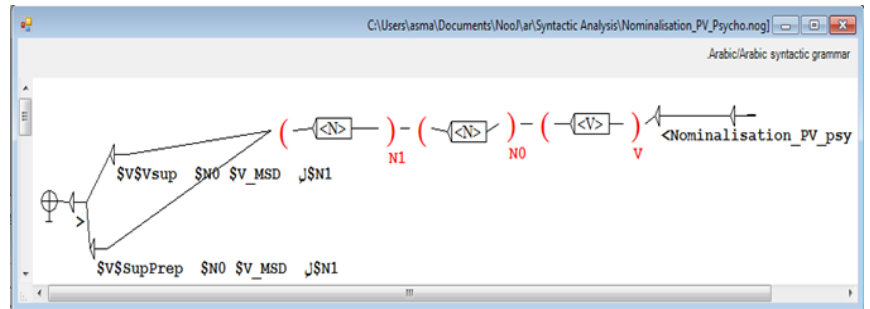


*Transformational
grammars
Arabic psychological
verbs
Journalistic texts
Syntactic-Semantic
analysis
NooJ*

References

Amzali, A. et al. "**Lexicon-Grammar Tables Development For Arabic Psychological Verbs**". In: Fehri H., Mesfar S., Silberstein M. (eds) *Formalizing Natural Languages With Nooj 2019 And Its Natural Language Processing Applications*. Communications In Computer and Information Science, Vol 1153. Springer, Cham, 2020

In this context, we will first list the transformations that can be applied to the Arabic sentences containing psychological verbs. After, we will implement a set of syntactic grammars modeling those transformations in Nooj platform, such as the nominalization grammar.



Formal Processing of Values of the Simple Kabyle Aorist Using NooJ

Hamid
Annouz

Abstract

Kabyle and Berber languages in general are languages with “aspect” which is defined as “a grammatical category that expresses the representation made by the talking subject of the process expressed by the verb ... i.e. the representation of its duration, its development or its completion...”.

In Kabyle, the verb expresses two “aspects”, an “accomplished” corresponding to the positive and negative “preterit” (= two verbal forms) and the “unaccomplished” corresponding to the simple and intensive “aorist” (= two verbal forms).

The first describes a completed action and the second an unfinished action. As for the temporal value and the aspectual values (inchoative, imminent, progressive, etc.), grouped in these two categories (accomplished / unaccomplished), they can only be defined by the context. More precisely, we must combine together the morphological information that we have about the verb, but also the semantic nature and the grammatical construction of the sentence.

Involving the NooJ syntax grammars, we will try here to describe and formalize the different theme values of the simple “aorist”. We can divide the simple “aorist” this way:

- aorist without a particle: expresses the injunctive, the contingency, the optative chained and the sequence of the process;
- aorist with a particle (ad): expresses the future, the optative (positive / negative) and the oath.



INALCO
[Paris, France]



hamid.annouz
@gmail.com



Verb values
Formalization
Syntactic grammars
Kabyle language
NooJ

References

Chaker S., “**Aspect**”, In :
Encyclopédie Berbère VII,
EDISUD, Aix-en-Provence, 1989,
Pp.971- 977.

Galand L., **Regard sur le berbère**,
2ème éd. Achab, Tizi-Ouzou,
2013.

Nait-Zerrad K., **Grammaire
moderne du kabyle, tajerrumt n
teqbaylit**, Karthala, Paris, 2001.

Selberztein M. **La formalisation
des langues, l’approche de NooJ**,
ISTE Editions, Grande-Bretagne,
2015.

Formalizing Arabic Deverbal Noun 'Gerund' Using NooJ Platform

Ilham
Blanchete

Mohammed
Mourchid

Samir
Mbarki

Abdelaziz
Mouloudi

Abstract

The paper presents a linguistic study of the deverbal noun "Gerunds", also known as "al-masdar-المصدر" in the Standard Arabic language. Due to the irregularity of this category, a set of rules and conditions have been extracted to implement a comprehensive linguistic resource, which serve to enhance the linguistic analysis of Arabic texts. This resource has been implemented to complete our previous work, which is "formalizing standard Arabic resources using root and pattern approach", this resource will be implemented using the NooJ platform.

Arabic al-Masders are typically a mixed category, they share morpho-syntactic and semantic features with verbs, like: root, meaning class. "Al-Masders" also share the same morphological nature as nouns, like: the number and gender properties. We have implemented this resource using a pre-implemented verb resource, we have extracted conditions from the latter to restrict the generation process of gerund's models, each Arabic verb model has one or more gerunds, e.g. the verb (to write – KaTaBa-كَتَبَ) that has the root (KTB-ك ت ب) is linked to the gerunds model: (KaToB-كُتِبَ), (KiTaAB-كِتَاب) and (KiTaABaAP-كِتَابَة). These previous gerunds share the same morpho-syntactic and semantic features with the verb (to write – KaTaBa-كَتَبَ).

Rules have been also extracted from standard Arabic grammar books:

- Morphological condition, if a verb pattern is (AaFoEaLa-أَفْعَلَ) then its gerund's pattern is (liFoEaAL-إِفْعَال).
- Syntactic condition, if the pattern of a transitive verb is (FaEaLa-فَعَّلَ) or (FaEiLa-فَعَّلَ) then its gerund pattern is (FaEoL-فَعَّلَ).
- Semantic condition, if the verb is classified as a verb of emotion, and if it also meets other morphological features, then it must

MIC research team
Laboratory
MISC Ibn Tofail
University Kenitra
[Kenitra, Morocco]

ilham.blanchete@gmail.com
mourchidm@hotmail.com
mbarkisamir@hotmail.com
mouloudi_aziz@hotmail.com

Deverbal nouns
Gerund
Al-Masdar
Transformational grammars
Lexical analysis
Arabic language
NooJ

References

- Blanchete I., Mourchid M., Mbarki S., Mouloudi A. (2018) **Formalizing Arabic Inflectional and Derivational Verbs Based on Root and Pattern Approach Using NooJ Platform**. In: Mbarki S., Mourchid M., Silberztein M. (eds) Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications. NooJ 2017. Communications in Computer and Information Science, vol 811. Springer, Cham.
- Kermers J. (2007) **The formation of deverbal nouns in Arabic draft**. University of Frankfurt, Germany

be linked to the Mimi-gerund class, which is one of gerund types. The used verb resource has been implemented using a root-pattern approach, which helps to link each gerund with its verb model.

Regardless of the previous features that gerunds share with verbs, it has the same morphological nature as nouns:

- it can accept the definite article e.g. the gerund (the description – AL-WaSoF- الوصف) that has the root (WSF- وصف) and the pattern (FaEoL- فَعْل), shares the root and meaning class with the verb (to describe- WaSaFa- وَصَف) that has the conjugational model (FaEaLa-YaFoEiLu).
- Affixes can be added to extract the possible gerund's inflectional forms.
- Broken plural/ regular plural forms also can be generated from singular gerund forms, e.g. (descriptions -AWoSaAF- أَوْصَاف).

We have taken into account the previous morphological characteristics during the formalizing of the gerund resource.

In Arabic morphology, gerunds have both verbal and nominal use, it may appear in many positions in the sentence, it can take the place of a subject, object, we can even replace a gerund with a verb, E.g. A subjunctive case of the verb (to write – KaTaBa- كَتَبَ), which is (AaKoTuBa- أَكْتُبُ), can be replaced by the gerund (writing- KiTaABaAT- كِتَابَةٌ), the verbal sentence can be represented as the nominal sentence “يجب كتابة الدرس”, both of these two sentences have the same meaning, which is “I have to write the lesson”. The relationship between these two sentences has been implemented in NooJ platform as a transformational grammar, we have implemented syntactical grammars, which serve to replace the verb with the corresponding gerunds.

As a conclusion, we are going to implement a new Arabic resource, which serves to enhance the performance of linguistic analyzers using the NooJ platform, this resource will be linked to our implemented verb resource. Gerunds will be implemented using root-pattern approach, rules and conditions have been extracted from an implemented verb resource to restrict the generation process of gerunds, we are going to implement several transformational rules using the gerunds resource.

List of Contributors

A

Aladrović Slovaček	52
Alves	30
Amzali	92
Annouz	94
Aouini	62

B

Barahona	76
Barnes	68
Barreiro	18
Bartulović	60
Bekavac	30
Belhoucine	38
Ben Ghazela	36
Bénet	54
Bessaies	36
Bigey	74
Boulaalam	72
Bounoua	56

C

Capone	48
Caraty	62

Chadjipapa	64
------------------	----

D

Dardour	40
di Buono	16
Duran	88

E

El Hannach	56
Elia	58
Eriquez	48

F

Fehri	40
Fernández Gallino	80

G

Giulio	48
Greco	48

H

Haddar-----32, 40

I

Ilham Blanchete-----96

K

Kaigorodova-----20

Kasunić-----14

Kiseljak-----14

Kocijan-----26, 82

Kourtin-----42, 50

Koza-----76

Krek-----2

Kurolt-----26

L

La Ragione-----48

Landsman Vinković-----82

M

Machonis-----10

Maisto-----58

Manna-----16

Marcone-----48

Melillo-----58

Mesfar-----36

Mijić-----60

Mikelić Preradović-----68

Monteleone-----46

Monti-----16

Mota-----18

Mourchid-----38, 50, 92, 96

P

Papadopoulou-----64

Pascucci-----16

Pelosi-----58

Peti-Stantić-----4

Petrovski-----28

Piton-----12

Prihantoro-----70

R

Reyes-----80

Rhazi-----72

Rodrigo-----80

S

Saint-Joanis-----86

Santos-----18

Savarese-----48

Silberztein-----8

Š

Šojat-----26

T

Tadić-----30

Thakkar-----68

Torjmen-----32

W

Watanabe-----90

A

adjectives, 4, 18, 22, 26, 38, 72
adverbs, 4
ambiguity, 10, 32, 90
artificial intelligence, 46
Artificial Intelligence systems, 2
automatic opinion analysis, 20, 74

C

collocations, 2
computational analysis, 14, 76
computational stylometry, 16
concepts, 8, 16, 20, 38
 conceptual metaphor, 16
 relevance of a concept, 8
 universal concepts, 2
concreteness, 4
connectors, 80
 causal connectors, 80
 consecutive connectors, 80
 counterargumentative connectors, 80
corpora, 10, 18, 50
corpus, 2, 8, 16, 26, 28, 32, 38, 40, 42, 76, 80, 86
 1. to 4. grade students' literary works, 52
 1. to 4. grade students' textbooks, 52
 1. to 4. grade students' written works, 52
Arabic newspapers, 92
blog posts about dream descriptions, 16
CHILDES - children's language. *See*
classic and contemporary children fiction, 58

e-mails and social network communications, 74
Facebook and Twitter texts, 32
Learning Language Texts in Greek, 64
manually annotated corpus, 2
Medieval Latin wills, 60
Middle French texts from PALM, 62
Moroccan legal documents, 42
morphologically annotated, 70
PubMed database, 20
Russian National Corpus, 54
SFU, 68
stories for kids written by students, 80
student learner, 82
web corpus, 4
corpus linguistics, 8, 10

D

data modeling, 2
derivation, 8, 12, 54, 86, 90
diatheses, 90
dictionary, 16, 40, 56, 60, 64
 bilingual dictionary, 32
 dictionary matrix, 2
 digital dictionary, 2, 26
 e-dictionary, 10, 56, 58, 72, 76
 Erfan-DIC, 56
 historical dictionary of Arabic, 56
digital humanities, 8, 10, 14, 60, 62
digital intelligence, 48
disambiguation, 22, 70
 word sense disambiguation, 2

domain
French food, 22
journalism, 92
legal domain, 38, 42, 60
medical domain, 20, 26, 36, 40

E

entities, 8
ethos analysis, 74

F

figurative language, 16
finite transducer, 32, 46
fuzzy logic, 46

G

gerund, 96
grammars
disambiguation, 10
dissambiguation, 10
inflectional, 62
local grammars, 30, 36, 38, 56
morphological, 26, 28, 38, 60, 64, 70, 82
syntactic, 28, 32, 50, 70, 80, 82, 94
transformational, 72, 88, 92, 96

H

homographs, 8, 32
human-machine interfaces, 46

I

idioms, 10
imageability, 4
inflection, 8, 62
information extraction, 28, 36, 68
information retrieval, 8

L

language data, 2
language teaching, 80, 82
languages
19th century French, 12
American English, 10
Arabic, 32, 36, 38, 40, 42, 56, 72, 96
Belarus, 20
Berber, 94
Croatian, 4, 14, 26, 52
English, 10, 68
French, 8, 12, 22, 50

German, 82
Greek, 64
Indonesian, 70
Italian, 58
Kabylian, 94
Late Modern English, 10
Latin, 12
Macedonian, 28
Medieval Latin, 60
Middle French, 62
Old English, 10
Portuguese, 18, 30
Portuguese from Brazil, 18
Portuguese from Portugal, 18
Quechua, 88
Rromani, 90
Russian, 54
Slovene, 2
Spanish, 76, 80
Tunisian dialect, 32
Ukrainian, 86

lexical database, 2
lexical density, 52
lexical diversity, 52
lexical fields, 8
Lexicographic Data as a Service, 2
lexicography, 2
lexicon, 50
frequency lexicon, 58
inflectional lexicon, 4
lexicon of named entities, 28
lexicon of proper nouns, 28
lexicon of toponyms, 28
morphological lexicon, 28
linguistic data, 2, 54, 56, 64
cross-lingual, 2
monolingual, 2
linguistic engine, 8
linguistic units, 8
linked data, 2

M

morphological paradigms, 2
morphology, 12, 62, 64, 86, 94
morphosyntax, 90, 96
multiword expressions, 2, 8, 10, 22, 26, 64

N

named entity, 28, 30, 32, 36, 40
narrative analyses, 8
negation, 68
NLP, 2, 12, 20, 28, 42, 46, 48, 50, 80, 82
nominal ellipsis, 76

NooJ, 8, 10, 12, 14, 20, 22, 26, 28, 30, 32, 36, 38, 40, 42, 46, 48,
50, 52, 54, 56, 58, 60, 62, 64, 68, 70, 72, 74, 76, 80, 82, 88,
90, 92, 94, 96
nouns, 4, 8, 22, 26, 32, 38, 56, 60, 64, 72, 96

O

ontology, 38, 42, 46
orthography, 8, 28, 80

P

paraphrases, 72, 88
 emotion paraphrases, 18
particles, 10, 56
patterns, 2, 16, 36, 56, 96
pedagogy, 80, 86
pragmatics, 12
prepositions, 10, 38
psycholinguistic database, 4

Q

question answering systems, 36, 40, 42
 complex questions, 40
 definition question, 40
 factoid questions, 40

R

reduplications, 70
regular expressions, 68
relations, 8
repository
 concept repository, 2

S

semantic roles, 2
semantics, 54
semio-linguistic analysis, 74
sentiment analysis, 46, 68, 92
statistical analysis, 8, 48
syntactic analysis, 50, 72, 82, 88

T

taxonomies, 38
text analysis, 8
text mining, 20
translation, 32, 88

V

variants, 8, 62
verb, 4
verbs, 8, 10, 54, 56, 62, 86, 90, 94, 96
 auxiliary, 72
 non-pair verbs, 86
 phrasal verbs, 10
 psychological, 92
 sentiment verbs, 88
 support verbs, 18, 72
 transitive, 96
vocabulary
 children vocabulary, 58
 evolution of the vocabulary, 8
 preschool-children vocabulary, 52
 school-children vocabulary, 52

W

word class, 4