THE 18TH NOOJ INTERNATIONAL CONFERENCE, 2024

Book of Abstracts

Bergamo (Italy), June 5th to 7th, 2024

Editors

Agathe Semlali
X23 Science in Society, Italy
Treviglio, Italy

Cecilia Olivieri
X23 Science in Society, Italy
Treviglio, Italy

Lorenzo Maggio Laquidara
Duke Kunshan University, China [1]; X23 Science in Society, Italy [2]
Kunshan, Suzhou, China [1]
Treviglio, Italy [2]

Jiwon Shin
X23 Science in Society, Italy
Treviglio, Italy

Mario Monteleone
University of Salerno
Salerno, Italy

Max Silberztein
Université de Bourgogne Franche-Comté
Besançon, France

—

Publisher

X23 Science in Society **(www.x-23.org)**

Contents

Preface (IX)

INVITED TALKS 3

Automatic Translation and AI

Prof. Stefania M. Maci 4

MORPHOLOGICAL & LEXICAL RESOURCES 56

Exposing Diminutive and Pejorative Verbs in Croatian

Krešimir Šojat, Kristina Kocijan 7

Some Linguistic Reflections on NooJ Electronic Dictionaries and Italian Morphological Derivation routines

Mario Monteleone 9

Automatic Grammatical Disambiguation in Belarusian and Russian Legal Domain

Valery Varanovich, Mikita Suprunchuk, Yauheniya Zianouka, Yuras Hetsevich 11

Linguistic and Morphological Analysis of Persian Verbs: Utilizing the NooJ Platform

Marzieh Rabiei 13

Quantifying Code-Switching: Metrics and Analysis in Ukrainian-Russian Bilingual Dataset

Olha Kanishcheva, Maria Shvedova 14

Recognizing Verbs in Medieval Latin

Linda Mijić, Anita Bartulović 15

Formalizing Ukrainian language variants with NooJ

Olena Saint-Joanis 16

Formalizing abstract nouns with the suffix "-pen" in Rromani

Masako Watabe 18

Implementation of an Arabic Grammar Checker using NooJ

Rafik Kassmi, Samir Mbarki, Abdelaziz Mouloudi 19

Local Syntactic Grammars for Quechua Negation and Negative Sentences

Maximiliano Duran 21

Linguistic Analysis of Explicit Language in Croatian Song Lyrics: A Comparative Study using NooJ

Mila Bikić, Valerija Bočkaj 23

NooJ Contributions to Second Language Acquisition: Virtual Reality Impact on Interlanguage, New Learning and Teaching Perspectives

Edoardo Di Gennaro 25

Annotation of Qualitative Linguistic Features of Text Readability in Italian Institutional Textse

Nicola Cirillo, Alessandro Maisto, Daniela Vellutino 29

——

SYNTACTIC AND SEMANTIC RESOURCES 29

A Set of Disambiguating Grammars for webnooj

Max Silberztein 30

A Collocation Analysis of the Verbs from Lemma <LAHIR> in Indonesian Translation of the Holy Quran and its Potential Implications for the Construction of NooJ Dictionaries and Grammars of Indonesian

Prihantoro 31

Automatic Processing of Free Word Combinations Containing Quantitative Expressions in NooJ

Yauheniya Zianouka, David Latyshevich and Yuras Hetsevich 33

Conditional Clauses versus Concessive Clauses in Spanish: An Automatic Treatment with the NooJ Platform

Andrea Rodrigo, Silvia Reyes and Mariana González 35

Exploring metaliterary cues using NooJ: the case of "Les Amours de Psyché et de Cupidon" by Jean de La Fontaine

Fabio Meroni 36

"Who's done it? ": detecting he or she crime

Filip Lukeš and Filip Vuković, Anela Rački 38

Construction of a Translation Relative Clauses System with NooJ

Nizar Jarray, Héla Fehri and Hajer Cheikhrouhou 39

——

NLP SOFTWARE APPLICATIONS 40

Linguistic Analysis for Legal Natural Language Question Analysis in Question Answering System Development

Ismahane Kourtin and Samir Mbarki 41

NooJ Digital Text Automation Workshop Artificial Intelligence (AI): Data Processing Automation and Cybersecurity Technologies

Ritamaria Bucciarelli, Francesco Terrone, Samuela Franceschini, Andrea Rodrigo and Javier Enriquez 42

From the social network data to the document-oriented NoSQL Datawarehouse using NooJ platform

Amira Abbes, Héla Fehri and Ines Ben Messaoud 44

Nooj Grammars and Cybersecurity. Possible Future Uses and Scenarios

Beatrice Folino and Mario Monteleone 45

Beneath Sentiment Analysis: a comparison of polarity tagging approaches

Luigi Celardo, Giuseppe Giordano, Michelangelo Misuraca and Mario Monteleone 47

——

CORPUS LINGUISTICS & DISCOURSE ANALYSIS 49

Using NooJ in action research: the case of personal expression in professional situations

Magali Bigey 50

"Are you the man of culture?" Analysis of YouTube Comments on Male and Female Athletes

Bruno Križić and Patricija Petanjek 51

Drama of the Heart: A Journey into Emotional Dynamics in Anna Karenina

Lucija Bažant and Mate Grizelj 52

Climate Change and Migration: Government Officials' Narratives at COP26 and COP28

Jiwon Shin 53

Corpus-Based Exploration of Defense-Related Political Speeches

Maja Ivanovic and Kristina Kocijan 54

How (not) to Talk about Feminicide and Gender-based Violence: An Analysis of Narrative Strategies Adopted by Italian Newspapers

Lisa Pezzali 56

Victimization and otherness: Unearthing colonial and post-colonial tropes in the E.U. migration policy arena using NooJ

Davide Carminati, Lorenzo Maggio Laquidara 57

——

SOCIAL SCIENCES APPLICATIONS 58

Unveiling Migrant Entrepreneurial Narratives: A Discourse Analysis Approach for Inclusive Support

Cecilia Olivieri, Davide Carminati, Lorenzo Maggio Laquidara, and Agathe Semlali 59

Beyond Sustainable Development, Unravelling Regenerative Tourism Through NooJ.

Lorenzo Laquidara, Agathe Semlali 61

Syntactic Analysis of Media Reporting on Feminicides and Gender-based Violence in Modern Italy

Giorgia Bozzetto, Nicola Cannoniero, Caterina Lausi, Stefano Madonna, Camilla Nastro, and Anna Maria Petruzzi 63

Linguistic Analysis on Project Evaluation through NooJ: the EU WeMed_NaTour Pilot Trip Assessment

Agathe Semlali, Jiwon Shin, and Marika Mazzi Boém 65

PREFACE

When, between the end of May and the start of June 2023, from a balcony leaning on the Adriatic Sea at the University of Zadar, they asked us to curate NooJ's next year International Conference, we accepted without blinking an eye. Immediately, we got to work on it. Some years have passed since the NooJ Community gathered in Italy for their traditional meeting: the last time was in 2018, in Palermo, beyond the boot's tip in mid-May. In 2017 the Community visited our colleagues in Kenitra, Morocco. In 2016, the Czech Republic. The year before, the Academic of Sciences in Minsk, Belarus. The Conference had already taken place in Italy, exactly ten years ago: it was at the Department of Humanities and Social Sciences in Sassari, Sardinia (2014). Then, until today other host cities have included: the atheneum in Hammamet, Tunisia (2019); in Besançon, France (2021); in Rosario, Santa Fe, Argentina (2022). Plus, a virtual one in Zagreb, Croatia, in 2020: while the COVID-19 pandemic secluded the planet – blaming the human kind about that frivolous idea of an unrivalled Enlightenment *–just come in–* Now [Pinker, S., 2018].

We are X23, not a university.

We first encountered NooJ in 2012, while conducting some studies on theme-based bibliography classification methods: we had discovered that by modifying the code in NooJ, that's an open software, we could aport modifications onto a Dewey system, and install a novel loose semi-hierarchy [The Primate, 2012]. *Some time later, we challenged with a speech-synthesiser, a micro-recorder pasted on the skin tract, next to the epiglottis: an emotional language sensor – that for instance, could capture the frightened woman's inner speech at the moment of an aggression* [Against, 2013]. *Then, further, a cognitive machine* [GDC+, 2015]; *and a semantic assistant to aid description of complex objects: in a metropolis for instance, to decipher how cities think* [BeLive3D, 2016].

In our lab, we study the ways Sciences are found in Societies: we self-sustain, and work as an independent action-research body.

The examples just mentioned apply to Sociolinguistics. Hence, we hereby proposed to inaugurate a new topic in the Nooj Call for Papers, 2024, i.e. "NLP Societal Applications and Citizen Science". In so doing, the Scientific Committee recognised "Migrations", "Geography of Population Flows", "Counter-tourism and Territories Reparation", "Political Discourse", *and* "Gender Studies" *as new disciplines of interest to their scientific Community.*

Nevertheless, we had to sustain the NooJ Conference among fund scarcity. The University of Bergamo, we couldn't be more grateful for, came to our rescue: their Department of Foreign Languages, Literature and Cultures hosted us in their classrooms without asking for a euro. All the catering suppliers charged us low-cost factory price. Cecilia, Agathe, Lorenzo, Davide, Jiwon, and Marika put our current studies on community semantics, and the interviews cycles with the "23Connected Communities" *migrant members, to a temporary halt, to set up the scene for the forthcoming start of the 3-days peace conflict; for the debates, and the prolific contrast of ideas; towards those shared obsessions for a most profound understanding of phenomena that engulf us. A UV-infraction, in the history of the NooJ Conference. An exploration of the ancestor linguistic characters; an homage to the* ξένος (xènos) *origin of lexicon-grammar* [Gross., M. 1963~]. *A practical and formal sign of truce, silence, and concert. A good read to you.*

— X23 Science in Society (www-x-23.org)




Nooj International Association
Association Internationale des utilisateurs de Nooj


UNIVERSITAS STUDIORUM BERGOMENSIS

INVITED TALK

Automatic Translation and AI

Prof. Stefania M. Maci

Università degli Studi di Bergamo

Bergamo, Italy
stefania.maci@unibg.it

Abstract

The spread of AI applications has triggered scientific, philosophical, political, social, ethical and legal debates about the pros and cons of AI for humanity. On the one hand, AI is seen as complementary to the human mind [1]. On the other hand, AI is assumed to pose an existential risk to humanity. Both sides agree that AI must be enabled to provide the right assistance to humans by being properly trained and regulated to ensure coexistence between humans and this new revolutionary entity (cf. [2]). To be operational, AI uses large language models (LLMs) to execute instructions. LLMs are mathematical models that calculate either the probability that a sequence of words W_1 to W_n can occur together in a sentence [3][4], or the similarity between word sequences [5], thanks to statistics such as likelihood. Language modelling is primarily associated with Natural Language Processing (NLP), which has recently been revolutionised by a new LLM architecture called *Transformer*. This is a type of deep learning modelling architecture that uses multiple layers of nodes or "neurons" [6] to learn complex patterns and relationships from data. *Transformer* was first introduced by Vaswani et al. [7] for machine translation. In the case of translation, the *Transformer* takes as input the sequence W_n (such as "I like my place because it is cosy") in the encoder and the already translated sequence W_{n-1} ("adoro la mia casa perché è") in the decoder to output the word ("cosy") that has the highest probability of being correct in the target language (i.e. "accogliente").

With the advent of AI conversational interfaces and chatbots, the debate has mainly focused on the 'ability' of AI to translate an infinite variety of texts, including specialised texts. Recent developments in AI have shown that effective chatbots can be built on the basis of an LLM that has been trained with vast amounts of human language, mainly from books and the internet. Their main function is to assist users in generating coherent, contextually relevant and informative text by understanding and predicting words, sentences and even paragraphs. Hence, the aim of this paper is to investigate how chatbots "translate" texts. Specifically, I will try to identify similarities and differences between different chatbots. In doing so, I will also examine the extent to which the translations produced by chatbots may differ from the professional patterns of human translations.

As claimed by Monteleone (2022) [8], Neural Machine Learning (NML), fails to guarantee constant satisfactory results, which, on the contrary, appear to be possible with NooJ [9]. Indeed, "NooJ grammars can achieve efficient and effective machine translation, also demonstrating that no correct translation procedures is achievable without formalized morphosyntactic rules about contexts, co-occurrence and selection restriction" [8].

Keywords

AI, Chatbot, Translation, NLP (Natural Language Processing), NooJ

References

1. Hughes, Alex (2023). Why AI will ultimately lose the war of creativity with humanity. BBC Science Focus, https://www.sciencefocus.com/future-technology/ai-art-creativity-war (15th March 2024)
2. EU Parliament, (2018, 2021) Coordinated Plan on Artificial Intelligence. Available at: https://digital-strategy.ec.europa.eu/en/policies/plan-ai;
3. Manning, Chris D., and Schütze, Hinrich (1999). Foundations of statistical natural language processing. MIT press
4. Jurafsky, Daniel and Martin, James H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics, and Speech Recognition. Prentice Hall PTR, USA. Available at: https://web.stanford.edu/~jurafsky/slp3/
5. Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, and Jauvin, Christian. (2003). A neural probabilistic language model. Journal of machine learning research 3: 1137-1155.
6. McCulloch, Warren S., and Pitts, Walter. (1943) "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical Biophysics, 5(1-2), 115-133.
7. Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review (15th March 2024)
8. Monteleone, M. (2022). Zellig S. Harris' Transfer Grammar and Its Application with NooJ. In: González, M., Reyes, S.S., Rodrigo, A., Silberztein, M. (eds) Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2022. Communications in Computer and Information Science, vol 1758. Springer, Cham.
9. Silberztein, M.(2016) Formalizing Natural Languages: The NooJ Approach. ISTE Ltd and Wiley, London

MORPHOLOGICAL & LEXICAL RESOURCES

Exposing Diminutive and Pejorative Verbs in Croatian

Krešimir Šojat

Faculty of Humanities and Social Sciences

University of Zagreb
Zagreb, Croatia
ksojat@ffzg.hr

Kristina Kocijan

Faculty of Humanities and Social Sciences

University of Zagreb
Zagreb, Croatia
krkocijan@ffzg.hr

Abstract

This paper addresses the recognition of diminutive and pejorative verbs in the Croatian language using NooJ. Diminutiveness is a category characteristic for different types of words, primarily for nouns, verbs and adjectives. Descriptions of diminutives can be focused on (a) their morphological characteristics, (b) the specifics of the syntactic constructions in which they are found, or on (c) their semantic properties. In Croatian grammars and reference books, verbal diminutiveness is approached as part of the verbal Aktionsart and as a phenomenon of derivational morphology. Pejorative verbs are not tackled as part of Aktionsart descriptions, but they are usually listed together with diminutive verbs in derivational descriptions due to very similar word-formation processes by which they are derived from verbal stems.

The article deals with the word-formation morphology of Croatian verbs, more precisely with derivation. Derivation includes various processes, such as prefixation, suffixation, simultaneous prefixation and suffixation, ablaut (usually combined with some kind of affixation), etc. Suffixation is the most frequent process in the formation of Croatian words. Diminutive and pejorative verbs in Croatian are formed by prefixing and suffixing. In works that deal with the derivation of such verbs, only two prefixes are cited that add a diminutive meaning to the formative base (e.g. *gristi* 'to bite – *prigristi* 'to have a (small) bite/nibble'). On the other hand, there are approximately twenty suffixes used to derive verbs with diminutive and pejorative meanings from other verbs. In this paper, we deal exclusively with suffixal formation of verbs.

In terms of morphological structure, Croatian verbs usually contain two suffixes. In verbal lemmas (i.e. in the infinitive forms), one suffix is used to indicate the verbal aspect and conjugational type, and the other to indicate the infinitive (e.g. *pisati* 'to write' – *pis-a-ti*). The specificity of diminutive and pejorative verbs is that they contain an additional suffix, that is, their suffixal structure consists of three suffixes. A suffix indicating diminutiveness or pejorativeness is always found between the verb root and the remaining two suffixes (e.g. *pisati* 'to write' – *pisuckati* (*pis-**uck**-a-ti*) 'to write a little bit; *piskarati* (*pis-**kar**-a-ti*) 'to scribble'). Based on such a regular morphological structure, the morphological grammar in NooJ is structured to recognize diminutive and pejorative verbs.

The objectives of the work are as follows:

1. to create a morphological grammar that is capable of recognizing:

    1.a. verbs that contain a diminutive or pejorative suffix in their suffixal structure;

    1.b. verb stems used for the derivation of verbs with diminutive and pejorative meanings;

2. to apply the grammars on a large corpus and automatically extract all the occurrences;

3. to sort out extracted verbs and add the new entries to the NooJ dictionary for the Croatian language, i.e. verbs with diminutive and pejorative meanings. In its current form, the NooJ dictionary for Croatian contains a very small number of such verbs.

The results will be evaluated manually. The obtained results are useful because they contribute to the development of resources for the Croatian language within the NooJ platform. Although diminutive and pejorative suffixes are listed in the literature, a complete list of verbs expressing such meanings does not exist. Therefore, the results can be used for further research on formative morphology in the Croatian language. The results obtained from this research are also useful because they represent further steps towards creating an enhanced morphological parser for the Croatian language.

Keywords

Diminutive verbs, Pejorative verbs, Morphology, Suffixation, Croation, NooJ

Some Linguistic Reflections on NooJ Electronic Dictionaries and Italian Morphological Derivation routines

Mario Monteleone

Università degli Studi di Salerno

Fisciano, Salerno, Italy
mmonteleone@unisa.it

Abstract

As known, NooJ Derivational Routines [1,2] are especially useful for speeding up the creation of those exhaustive electronic dictionaries necessary to achieve complete and in-depth NLP analyses. Furthermore, we base the structuring of these routines on precise morphosyntactic and derivational rules, such as, for example, in Italian, the one that predict, for each transitive verb, an adjectival derivation in –bile, as shown below:

Max riscrive il documento =: Il documento è riscrivibile da Max
(Max rewrites the document =: *?The document is rewritable by Max)

However, we must not forget that the Italian lexicon - together with many other lexicons - is not structured: this means that, as we will see, the set of words that it includes, in their morphological form, does not allow us to produce all the processes of meaning - grammatical and acceptable - in the same way. As for this mismatch, the philological derivative aspect of the words themselves also plays an important role. It therefore follows that a wildcard application of the derivation routines, although well-structured thanks to solid and validated morphosyntactic rules, can produce imprecise results, in terms of the absence of both specific entries and specific meanings and senses. Ultimately, such errors can invalidate the results of many different NLP analyses.

For instance, many Italian iterative verbs, obtained with the prefix -ri placed at the beginning of specific infinitive verbs, should not be used as electronic dictionary entries, as their occurrences do not produce simple sentences. In fact, they can be transformed into sentences containing metadiscursive (hence, erasable) adverbs:

Max riscrive la sceneggiatura = Max scrive la sceneggiatura di nuovo
(Max rewrites the script = Max writes the script again)

Max riracconta la barzelletta =: Max racconta la barzelletta nuovamente
(Max retells the joke =: Max tells the joke again)

To this, it must be added that in Italian there are verbs which are only apparently iterative. This is the case with verbs suc as rilevare (to detect), which does not only mean levare di nuovo (to take off again), rifutare (to refuse), which does not only mean fiutare di nuovo (to smell again), or rispettare (to respect), which does not only mean spettare di nuovo (to be up again to).
Moreover, still regarding the aforementioned derivation of adjectives in -bile, it should be noted that some of them are derived from intransitive verbs, as passeggiabile (walkable) from passeggiare (to stroll).

Consequently, in our study, we will try to establish which NooJ derivational routines can be structured and applied, in Italian [3], via wildcard commands, that is when and how apply automatic morphological derivations based on word lexical uses and syntactic profiles. Other word classes to include in such analysis, in addition to those already

mentioned here, will be for instance those of the mode adverbs in –mente, as milanesemente (in the Milanese way) derived from milanese.

Keywords

NooJ, NooJ grammars, NooJ derivational routines, NooJ finite-state automata and transducers, Italian lexicon and syntax, Italian derivational rules

References

1. Silberztein, M. (2003). The NooJ Manual. NooJ Association Web Page, http://www.nooj-association.org last accessed 2023/19/03.
2. Silberztein, M. Formalizing Natural Languages. (2016). The NooJ Approach. ISTE Ltd and Wiley, London.
3. Grossmann, M., Rainer F. (eds.). (2004). La formazione delle parole in italiano. Max Niemeyer Verlag, Tübingen.

Automatic Grammatical Disambiguation in Belarusian and Russian Legal Domain

Valery Varanovich

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

Mikita Suprunchuk

United Institute of Informatics Problems
Minsk, Belarus
suprunchuk@gmail.com

Yauheniya Zianouka

United Institute of Informatics Problems
Minsk, Belarus
evgeniakacan@gmail.com

Yuras Hetsevich

United Institute of Informatics Problems
Minsk, Belarus
yuras.hetsevich@gmail.com

Abstract

The homonymy of language units at different levels is one of the most intractable problems in natural language processing [2; 4]. In 2023, the Belarusian NooJ team analyzed the homonymy of various types in the Belarusian language on the basis of composed legal domain corpus in NooJ [3]. Currently, we focus on one type – grammatical homonymy (coincidence of different word forms). It is a very common phenomenon in Slavic languages, including Belarusian [1]. For example, the forms of nominative and accusative cases of inanimate masculine nouns (край 'edge'), forms of genitive, dative, instrumental, and prepositional cases of feminine adjectives (маленькай 'small') and many others regularly match.

Recognizing grammatical homonymy is a core task for creating machine translation systems. Even when translating closely related languages, the case forms may vary: for example, case changes in the expression загадчык кафедры 'head of the department' while translating from Belarusian to Russian: the form of the genitive case is homonymous with the dative case, – заведующий кафедрой (the instrumental case).

In text-to-speech systems, the problem of stress arises in languages with non-fixed and mobile stress, including Belarusian. There are forms within the paradigm that differ only in stress: рукі [ru'ki] 'hand' is the genitive singular, and рукі ['ruki] is the nominative plural. Therefore, it is crucially important to identify which form of word is used in a particular sentence.

Previous research [3] revealed that grammatical homonymy is the most common among other types of homonymy in legal texts. This type covers about 25% of all cases, but it is already 78% in the total number of different words (lexemes). Within this work, we plan to use the full corpus of codes and conduct a detailed study, finally clarify the degree of prevalence of grammatical homonymy in Belarusian legal texts, and create NooJ grammars to solve grammatical homonymy.

Key words

Automatic text processing, Disambiguation, Grammatical homonymy, Legal texts corpus, NooJ

References

1. Barabash, O. V. (2015). Differentiation of homonymy and polysemy of legal terms (= in Russian: Барабаш, О. В.: Разграничение омонимии и полисемии юридических терминов). Rhema.=Рема (2), 39–51. https://cyberleninka.ru/article/n/razgranichenie-omonimii-i-polisemii-yuridicheskih-terminov, last accessed 2024/01/03.
2. Rakhilina, E. & Kobritsov, B. (2006). Semantic ambiguity as an application-oriented problem: word class tagging in the RNC (= in Russian: Рахилина, Е. В., Кобрицов, Б. П. Многозначность как прикладная проблема: лексико-семантическая разметка в Национальном корпусе русского языка). In: Laufier, N. (ed.) Kompjutiernaja lingvistika i intielliektualnyje tiekhnologii «Dialog 2006» (= Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции «Диалог 2006» / под ред. Н. Лауфер), pp. 445–460. RGGU Publisher, Moscow.
3. Varanovich, V., Suprunchuk, M., Zianouka, Y., Dolgova, A. & Hetsevich, Y. (2023). Creation of a legal domain corpus for the Belarusian module in NooJ: texts, dictionaries, grammars. In: Rodrigo, M., Silberztein, M. (eds.) 16th International Conference NooJ 2022, pp. 151–162. Springer Cham, Rosario (2023).
4. Zalizniak, A. A. (2006). Ambiguity and ways of representing it. Jazyki slavianskikh kultur, Moscow. (= in Russian: Зализняк, А. А.: Многозначность и способы ее представления. Языки славянских культур, Москва.)

Linguistic and Morphological Analysis of Persian Verbs: Utilizing the NooJ Platform

Marzieh Rabiei

Université Bourgogne – Franche-Comté

Besançon, France
marzieh.rabiei_karahroudi@univ-fcomte.fr

Abstract

This study presents a formalized description of the Persian language, commonly known as Persian or Farsi, leveraging the NooJ platform. The analysis encompasses various linguistic aspects such as vocabulary, spelling, inflectional, derivational, and agglutinative morphology, local and constituent syntax, along with transformational grammar. The endeavor underscores the necessity for a robust foundation in Persian linguistics, especially considering the absence of formalization for the French-Persian pair despite previous efforts in formalizing French for multiple languages. With over 100 million speakers worldwide, the imperative to advance research in this domain is indisputable.

Specifically, this research delves into the linguistic and morphological examination of persian verbs, employing the NooJ platform as its primary instrument. The project unfolds through several phases, starting with the gathering and curation of linguistic resources like French-Persian dictionaries, followed by the evaluation of morphological analyses applied to Persian corpora.

The first phase involves data collection, notably textual corpora, followed by the inventory of Persian vocabulary, with an emphasis on specific verb categories and distributional classes.

In the second phase, the morphological analysis of verbs is conducted. This includes describing conjugation and formalizing verbs, as well as establishing conjugation rules using the NooJ platform. Additionally, special attention is given to verb inflection, identifying lexical roots, affixes, and inflection rules.

Finally, the morphological descriptions are evaluated by applying these analyses to Persian corpora, thus verifying their validity and applicability in a real linguistic context.

This project aims to deepen our understanding of the linguistic structure of verbs in Persian and to demonstrate the effectiveness of using the NooJ platform in language morphological analysis.

Keywords

Persian language, NooJ platform, Linguistic analysis, Morphology, Syntax, Linguistic structure, Language research

References

1. Silberztein, M. (2015). La formalisation des langues : l'approche NooJ. ISTE.
2. Jahanshiri, A. (2023, November 15). Conjugueur de verbes persans. https://www.jahanshiri.ir/pvc/fr/
3. Karimi, S. (1997). Persian Complex Verbs: Idiomatic or Compositional? Lexicology 3(1). 273–318
4. Hafezian, M. (2001, October). LES VERBES PERSANS: Morphologie et conjugaisons. L'Harmattan.
5. https://github.com/reza-mashayekhi
6. https://www.iranketab.ir/book/84448-morphology-in-farsi

Quantifying Code-Switching: Metrics and Analysis in Ukrainian-Russian Bilingual Dataset

Olha Kanishcheva

University of Jena (1); SET University (2)

(1) Jena, Germany; (2) Kyiv, Ukraine
kanichshevaolga@gmail.com

Maria Shvedova

University of Jena (1); National Technical University - Kharkiv Polytechnic Institute (2)

(1) Jena, Germany; (2) Kharkiv, Ukraine
mariia.shvedova@khpi.edu.ua

Abstract

In the dynamic landscape of natural language processing (NLP), the analysis of code-switching, where speakers seamlessly blend multiple languages within a single discourse, poses a unique set of challenges [1]. The Ukrainian linguistic context, rich in cultural diversity and historical influences, is a fascinating domain for exploring code-switching phenomena. This research delves into the intricate task of language identification within Ukrainian code-switching corpora, shedding light on the complexities inherent in deciphering linguistic boundaries in a multilingual environment. This paper aims to explore the peculiarities of code-switching in Ukrainian datasets, highlighting the complexities in processing texts where Ukrainian and Russian are present and Ukrainian-Russian mixed speech (Surzhik), which contains hybridization within a word. The work also applies code-switching metrics to our data and finds that they quantify the frequency, patterns, and types of language alternation within linguistic data, shedding light on bilingual communication dynamics and language use patterns. These metrics provide valuable insights into the complexity and structure of code-switching phenomena in multilingual contexts, on the example of the Code-Switching Parliamentary Corpus. In our study, we selected sentences from the transcripts of Ukrainian parliamentary sessions [2] that demonstrate switching between Ukrainian and Russian and mixing the two languages.

In the original transcripts from 1990 to 2017, there are cases of switching from Ukrainian to Russian and vice versa [3]. All sentences were tokenized, and the language of each token was manually labeled. Categorization involved the allocation of tokens into five distinct classes: Ukrainian, Russian, Ukrainian-Russian hybridized words, Others, and Punctuation. The corpus is over 150,000 tokens in size. As part of our research, we not only created a code-switching labeled dataset but also analyzed the data obtained. We also proposed a model for language identification at the token level.

Keywords

Code-switching, dataset, Ukrainian, Language identification

References

1. Barik, A. M., Mahendra, R., & Adriani, M. (2019). Normalization of Indonesian-English code-mixed Twitter data. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, China, Association for Computational Linguistics, 417-424.
2. https://www.rada.gov.ua/documents/Stenbul_pz
3. Kanishcheva, O., Kovalova, T., Shvedova, M., & von Waldenfels, R. (2023). The Parliamentary Code-Switching Corpus: Bilingualism in the Ukrainian Parliament in the 1990s-2020s. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Dubrovnik, Croatia. Association for Computational Linguistics, 79-90.

Recognizing Verbs in Medieval Latin

Linda Mijić
Department of Classical Philology
University of Zadar
Zadar, Croatia
lmijic@unizd.hr

Anita Bartulović
Department of Classical Philology
University of Zadar
Zadar, Croatia
abartulo@unizd.hr

Abstract

In this paper, we present NooJ dictionaries, inflectional grammars for recognizing different types of Latin verbs (regular, irregular, defective, deponent, semideponent, impersonal), and morphological grammars for recognizing compound verbs. These resources complete the work on NooJ morphological resources for Medieval Latin.

We apply the compiled grammars to a corpus of last wills and testaments written in Medieval Latin in the Zadar commune (MedTest) [1]. Since the corpus comprises only one type of legal document, a limited number of verbs exist. However, the grammars show a high complexity even in this sample. Namely, the Latin verbal system is highly inflected [e.g., see 2]. Its finite forms distinguish three persons, two numbers, three moods (indicative, subjunctive, imperative), six tenses (present, imperfect, future, perfect, pluperfect, future perfect), and two voices. The Latin verbs are associated with three stems (the present, the perfect, the participle) to which endings are added: e.g., 3rd person plural *habe-nt – habu-erunt – habit-i sunt* (< *habeo, 2. habui, habitum*); *vide-nt – vid-erunt – vis-i sunt* (< *video, 2. vidi, visum*); *fer-unt – tul-erunt – lat-i sunt* (< *fero, ferre, tuli, latum*). The non-finite verbal forms have five nomina, i.e., verbal adjectives (participles, gerundives) and verbal nouns (infinitives, gerunds, supins). In dealing with Medieval Latin, special attention is also paid to the local peculiarities considering the various changes and varieties in verb formation [see 3].

Taking the above into consideration, there are two aims of this paper. The first aim is to recognize all verbal forms, i.e., for NooJ to connect a particular verb with its inflectional paradigm. The second aim is, although derivations can be entered in the NooJ dictionary, to investigate whether compound verbs can be successfully recognized using NooJ grammars (for instance, verbs derived by prefix *de-* such as *decedo* (< *de* + *cedo*), *deficio* (< *de* + *facio*), *dimitto* (< *de* + *mitto*), etc.).

Keywords

Medieval Latin, Zadar commune, Last wills and testaments, NooJ

References

1. Mijić, L., & Bartulović, A. (2021). Formalizing Latin: An Example of Medieval Latin Wills. In: Bekavac, B., Kocijan, K., Silberztein, M. Šojat, K. (eds.), Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities. 14th International Conference, NooJ 2020, Zagreb, Croatia, 5–7 June 2020, Revised Selected Papers, pp. 24–36. Springer, Cham.
2. Moorwood, J. (1999). Latin Grammar. Oxford University Press.
3. Stotz, P. (2002). Handbuch zur lateinischen Sprache des Mittelalters. Bedeutungswandel und Wortbildung (2), Verlag C. H. Beck, München.

Formalizing Ukrainian language variants with NooJ

Olena Saint-Joanis

CRIT (1), CREE (2)

(1) Besançon, France; (2) Paris, France
olena.saintjoanis@gmail.com

Abstract

Modern Ukrainian is the result of several literary variants, which developed in the early 19th century throughout the territory of today's Ukraine. Each variant included several dialects and was influenced by one or sometimes several dominant languages (Russian, Polish, German, Hungarian). It's worth noting that in the 18th and 19th centuries, the Ukrainian language was either banned from the public sphere or forbidden, so the few published texts were perceived as linguistic models[5].

In this context, standardization of the language and its spelling was not a simple task, and this is why so many orthographic standardization attempts were made. These included "zhelekhivka" (1886) and others.

During the Soviet period, 18th and 19th century literary texts were rewritten according to current orthographic standards, but with the emergence of online libraries, original-language texts have become accessible to the public and are attracting a lot of interest[3].

So we decided to start preparing an "Old Ukrainian" module to complement the normative Ukrainian module, and we'd like to set out our work in this paper.

NooJ allows us to describe these variants very flexibly and link them to normative lemmas[1,2,4,6,7].

First, we prepare a dictionary, in which we link a number of old variants to modern norms in the following way:

лїс,ліс,NOUN+Masculine+Common+Inanimate+FLX=ЛИСТ+"zhelekhivka"

We are also preparing some paradigms and morphological grammars specific to these ancient variants. For example, in "zhelekhivka", to form diminutive nouns and adjectives, the suffixes -ейк- and -ойк- were used instead of the suffixes -еньк- and -оньк-, i.e. малейкий instead of маленький [small]. In the standard Ukrainian module, these ALUs were described as derived forms using a morphological grammar. So we modify this grammar to recognize the old forms.

This module complements the Ukrainian module, allowing old texts to be annotated with morphological information and thus opening up the possibility of studying them.

Keywords

NooJ, Old Ukrainian, Zhelekhivka, Ancient variants

References

1. Bekavac, B., Kocijan, K., & Tadić, M. (2015). Near Language Identification Using NooJ. In *Formalising Natural Languages with NooJ 2014. Monti J., Silberztein M., Monteleone M., Di Buono M*, Cambridge Scholars Publishing.

2. Chevalier, G., Kasparian, S., & Silberztein, M. (2004). Éléments de solution pour le traitement automatique d'un français oral régional. Dans *TAL [Traitement automatique des langues]*, vol. 45, n°2, 41-62.

3. Holoiukh, L. V. (2021). Istoriia ukrainskoi literaturnoi movy. [History of the Ukrainian literary language]. Nadstyria, Lutsk.

4. Kasparian, S. (2008). The Acadian NooJ module: Automatic Processing of a regional oral. In *Linguistica Atlantica,* n°29, 117-136.

5. Nimchuk, V., Puryayeva, N. (2004). Історія українського правопису: XVI–XX століття: хрестоматія [History of Ukrainian spelling: XVI-XX centuries: a textbook]. Naukova Dumka, Kyiv.

6. Pignot, H., Lardy, M. (2013). Formalizing a Dictionary of 17th Century English with NooJ. In *Formalising Natural Languages with NooJ : Selected Papers from the NooJ 2012 International Conference (Paris, France). Edited by Anaïd Donabédian, Victoria Khurshudian and Max Silberztein,* Cambridge Scholars Publishing, Newcastle, 37-47.

7. Watabe, M. (2018). A NooJ dictionary for the rromani language: toward a NooJ-relevant sorting of morphosyntactic tags. In *Formalising Natural Languages with NooJ and Its Natural Language Processing Applications. Edited by Mbarki S., Mourchid M., Silberztein M.,* Springer CCIS Series #811, 3-15.

Formalizing abstract nouns with the suffix "-pen" in Rromani

Masako Watabe

CRIT
Université de Franche-Comté
masako.watabe@univ-fcomte.fr

Abstract

This article aims to formalize a derivational grammar of abstract nouns with the suffix "-pen" in Rromani using the NooJ development environment.

The suffix "-pen" is added to various types of word: mostly verbs, adjectives, nouns, and some adverbs, numerals, and prepositions. For example, *phirel* [to walk] > *phiripen* [walk], *sasto* [healthy] > *sastipen* [health], *amal* [friend] > *amalipen* [friendship], *śaj* [can, may] > *śaipen* [possibility], *jekh* [one] > *jekhipen* [unity], *karing* [toward] > *karingipen* [direction].

There are several dialectal variants ("-pen," "-pe," "-ben," "-po," and "-mos,") but the morphologies, both derivational and inflectional, are common[1][2].

What is the best way to formalize these phenomena? There may be three options: 1 writing a derivational grammar with the "DRV" property (".nof" file), 2 writing a productive morphology (".nom" file), and 3 adding derivatives as lexical entries (".nod" file).

The suffix "-pen" can be added to theoretically any adjective without semantic modification. Then, we could adopt a productive morphology to formalize this derivation. Regarding verbs, most derivatives with "-pen" do not change semantic value. However, we should not adopt a productive morphology because the suffix "-pen" does not apply to certain verbs. For example, the verb *lośal* [to be pleaded] is a derivative of the noun *loś* [joy], and the derivative with the suffix "-pen" of this verb does not exist. Therefore, we could adopt a derivational grammar for verbal derivatives. Regarding nouns, since derivatives with "-pen" have additional semantic values, we should treat them as lexical entries. In addition, verbal and adjectival derivatives which have specific semantic values (e.g., *marel* [to beat] > *maripen* [blow, fight, war], *bodlo* [blue] > *bodlipen* [sky, heavens]) should be added in a dictionary too.

Keywords

Derivational morphology, NooJ, Rromani language

References

1. Courthiade, M. et al.: (2009). *Morri angluni rromane ćhibăqi evroputni lavustik*. Romano Kher, Budapest.
2. Sarău, G. (2004). *Rromani Leksikologìa*. Editura Universității din București, Bucarest.

Implementation of an Arabic Grammar Checker using NooJ

Rafik Kassmi

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

Samir Mbarki

Ibn Tofail University

Kénitra, Morocco
mbarkisamir@hotmail.com

Abdelaziz Mouloudi

Ibn Tofail University

Kénitra, Morocco
mouloudi_aziz@hotmail.com

Abstract

Arabic is a Semitic language characterized by its intricate morphology and syntax. The multitude and intricate grammar rules of the language can be perplexing for the common user of a word processor. Arabic grammar is a highly intricate topic, with even native Arabic speakers lacking comprehensive knowledge of it. The presence of an elliptic personal pronoun, lengthy sentences, complex syntax, the omission of diacritics, and free word order all contribute to the difficulty of Arabic grammatical checking. The word processing technology exhibits substantial potential for growth, yet it frequently results in writing errors. Numerous widely-used software applications incorporate grammar checkers, either as an integrated feature or as a separate tool, with the purpose of identifying and correcting these errors.

Grammatical errors in written language can be classified into two categories: mechanical editing errors and cognitive errors. Mechanical editing errors arise when utilizing a word processor, stemming from actions such as cut-and-paste or insertion-deletion processes. To correct these errors, one might delete or substitute existing words. For example, missing the object of a verbal sentence, such as in the sentence كتب الدرس */kataba addarsa/* in place of كتب الطالب الدرس*/kataba altālibu addarsa/*, which means "the student wrote the course". Cognitive errors include errors that occurred when the writer did not know the correct spelling of the word, forgot it, or misconceived it. For example, typing لاكن */lākin/* for لكن */lakin/*, which means but, adds the letter ا */ā/*.

In our previous research, we developed a spellchecker for the Arabic language using NooJ, which we named Al Mudaqiq [1]. The primary aim of the study outlined in this article is to expand our spellchecker by integrating a grammar checker for Arabic [2]. This will be achieved by exploiting the capabilities of the NooJ platform and using its command-line, noojapply [3]. This grammar checker will: first, detect all grammatical errors using the El-DicAr dictionary [4] combined with the improved morpho-syntactic grammar handling agglutination [5] in NooJ. Then, generate corrections and candidate suggestions in NooJ. Next, rank the candidates in descending order.

Keywords

Arabic language, Grammar checker, Spellchecker, Grammatical errors, NooJ

References

1. Kassmi R., Mbarki S. et Mouloudi A. (2024). Spelling Error Detection and Correction for Arabic Using NooJ. Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications, NooJ 2023.
2. Sibertztein M. (2015). La formalisation des langues, l'approche NooJ. ISTE Edition, London.
3. Silberztein M. (2006). NooJ' Manual. http://www.nooj4nlp.net/NooJ%20Manual.pdf.

4. Mesfar S. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard, thèse doctorale de l'université de Franche Comté, France.

5. Kassmi R., Mourchid M., Mouloudi A. et Mbarki S. (2018). Processing Agglutination with a Morpho-Syntactic Graph in NooJ. Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications, NooJ 2017, vol. 811, pp. 40-51. Springer.

Local Syntactic Grammars for Quechua Negation and Negative Sentences

Maximiliano Duran

C.R.I.T.

Université de Franche-Comté

Besançon, France
duran_maximiliano@yahoo.fr

Abstract

This work aims to formalize the Quechua morphosyntactic negation constructions and their evaluation when applied to analyzing positive or negative sentences and particularly, sentimental expressions. We will present some local NooJ graphical grammars, showing the results of their applications on our corpus.

We aboard the typological question if Quechua as a multiple-negation language *manam chusaqchu* [it's not nothing], and study the use of double-negative as equivalent to affirmative constructions *manam wasinaqchu* [lit. he's not someone who does not have a house].

The Quechua language has several types of negation: lexical *mana* [no], *mana munanchu mikuita* [he does not want to eat], morphological *kanchu* [there does not] *manam qullqi kanchu* [there's no money], syntactic *manaraqmi chayamunchu* [he hasn't yet arrived] and semantic *mikunmanñchu saksasqa kaspaqa* [If he were satisfied, he wouldn't it].

Following the ideas of Montelone et al. [4], we constructed seldom neg_grammars in order to debug some type of negative sentences.

A negative sentence employs the suffix *chu* [negation] and the adverb *mana* [no) preceding the predicate and present sentences such as *mana mikunchu* [he does not eat], which is opposed to *mikunmi* [he does eat]. Using morphosyntactic grammars, I try to make a detailed analysis of the remarkable property of this negation adverb *mana* [no] and after the construction of the corresponding grammar containing one adverb-suffix or combination of two or three of those suffixes, I present the resulting inflected forms of the 'no' adverb: *manaraq* [not yet], *manapaschá* [probably non], etc., which allow for the generation of many negative paraphrases.

Keywords

Quechua negation, Negative suffixation, Graphical negation grammar, Negative paraphrasing

References

1. Duran, M. (2017). Dictionnaire électronique français-quechua des verbes pour le TAL. Thèse Doctorale. Université de Franche-Comté. Mars 2017.

2. Duran, M. (2020). Transformation and Paraphrases for Quechua Sentiment Predicates. In Proceedings of the 14th Intenational Conference, NooJ 2020. Zagreb. Springer, Switzerland.
3. Guardia Mayorga, C. (1973). Gramatica Kechwa, Ediciones Los Andes. Lima Peru.
4. Montelone Mario and Mirto I. M. (2021). Nooj Grammars for Italian Negation System and Sentiment Analysis. Procedings 15th International Conference, NooJ 2021. Besançon, France.
5. Perroud, Pedro Clemente. (1970). Diccionario Castellano - Kechwa Dialecto de Ayacucho. Lima. Edición.
6. Silberztein, M. (2016). Language formalization: the NooJ Approach. Wiley Eds.

Linguistic Analysis of Explicit Language in Croatian Song Lyrics: A Comparative Study using NooJ

Mila Bikić

Faculty of Humanities and Social Sciences

Zagreb, Croatia
bikic.mila05@gmail.com

Valerija Bočkaj

Faculty of Humanities and Social Sciences

Zagreb, Croatia
vbockaj@gmail.com

Abstract

This research explores the usage of explicit language within Croatian song lyrics, employing NooJ, a linguistic development environment software and corpus processor. The corpus was gathered by automatically scraping the Croatian lyrics site https://cuspajz.com/ and manually selecting songs that fit into each genre. This initial corpus was subsequently cleaned to correct common orthographic errors found in Croatian online text. Specifically, the letter c is often used in place of diacritics č and ć, z and s instead of ž and š, respectfully, and đ is substituted with dj. The final corpus consists of 250 songs encompassing a diverse range of genres, including pop, rock, R&B, hip-hop, and country. Within each genre, 50 songs were selected, evenly distributed between male and female artists. Additionally, the songs were further categorized based on the decade of their release. The primary objective was to design a syntactical grammar capable of recognizing and annotating explicit words and phrases, encompassing verbal insults, sexually explicit language, and references to excrement.

The results of the analysis revealed hip-hop as the genre with the highest prevalence of explicit language usage, particularly prominent during the 2010s. Consistent with expectations, male hip-hop artists displayed a greater propensity for profanity compared to their female counterparts in this study.

To contextualize these findings, a comparative analysis was conducted against previous research focused on English song lyrics. This analysis uncovered hip-hop as the genre with the highest frequency of explicit language and indicated a progressive increase in the usage of explicit language over successive decades. Notably, while the 1950s corpus displayed a complete absence of profanity, a stark contrast was observed in the 2020s corpus. The parallel trends evident in both corpuses suggest a significant evolution in societal norms and linguistic expression over time, wherein explicit language is increasingly ingrained within our cultural identity.

Keywords

Linguistic analysis, Explicit lyrics, NooJ, Croatian music

References

1. Chomsky, N. (1957). Syntactic Structures. Mouton, The Hauge.
2. Gregurović, M. (2021). 'Psovke u hrvatskome jeziku', Master's thesis, Juraj Dobrila University of Pula, cited: 06.02.2024., https://urn.nsk.hr/urn:nbn:hr:137:938450.
3. Mitkov, R. (Ed.). (2004). The Oxford handbook of computational linguistics. Oxford University Press, Oxford.

4. O'Keeffe, A. (Ed.) & McCarthy, M. (Ed.). (2012). The Routledge handbook of corpus linguistics. Routledge, New York.
5. Padovan, F. (2018). 'Komunikacijska obilježja psovke u hrvatskome jeziku', Diplomski rad, University of Zagreb, Faculty of Humanities and Social Sciences, cited: 13.02.2024., https://urn.nsk.hr/urn:nbn:hr:131:676524.
6. Silberztein, M. (2016). Formalizing natural languages: the NooJ approach. Wiley-ISTE, London.

NooJ Contributions to Second Language Acquisition: Virtual Reality Impact on Interlanguage, New Learning and Teaching Perspectives

Edoardo Di Gennaro

University of Bergamo

Bergamo, Italy
edoardo.digennaro@protonmail.ch

Abstract

Currently, new technologies offer a wide range of possibilities to explore many fields in linguistics. In fact, the potentialities of many software have been highlighted and developed to investigate the interaction between AI and linguistics. Indeed, many software have been strongly influenced by linguistic research (i.e. Chat GPT 3 and 4). In this paper, Second Language Acquisition [1] and Virtual Reality interaction will be investigated through NooJ software [2]. Language competence can be better conceived in the comparison between standard and non-standard linguistic items that speakers of different languages acquire and produce, while being exposed to L2 inputs for a specific period [3]. During this period, non-native speakers build their knowledge concerning the second language by constituting a temporary system of strategies, intakes and linguistic competence known as interlanguage [4]. According to the latest studies in Embodied Cognition and Sensory-Motor System, there are remarkable evidence that language acquisition, at least in terms of memory storage and meaning retrieving, involves many parts of our brain [5]. It is to say that our linguistic competence is strictly related to the way we experience the world all around (learning by doing) [6]. According to SLA studies, the elaboration of all the hypotheses, the mistakes and the intake concerning the second language ensure our language competence to be more and more similar to that of native speakers'. In order to understand the ways to correctly implement VR software for L2 learning [7], it could be very useful to investigate non-native speakers' interlanguage development by using NooJ. It would be essential to have a wider and clearer perspective of how interlanguage works, demonstrating the existence of common mistake-items or strategies related to a precise learning method/period of acquisition, which could remarkably influence SLA research and VR teaching and learning environments [8].

Keywords

Second language acquisition studies, NooJ, Virtual reality

References

1. Krashen, S. (1982). Principles and Practice in Second Language Acquisition. Pergamon Press, Los Angeles: University of Southern California.
2. Rodrigo, A., & Bonino, R. (2019). Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua. Rosario: Ciudad Gótica.
3. Balboni, P. (2008). Linguistica acquisizionale e glottodidattica. In Grassi R., Bozzone Costa R., Ghezzi C. Dagli studi sulle sequenze di acquisizione alla classe di italiano L2. Perugia: Guerra, 23-34.
4. Pallotti, G. (1999). Acquisire e insegnare l'italiano: dai processi naturali agli interventi didattici. In Favaro G. (a cura di). Imparare l'italiano, imparare in italiano. Guerini: Milano, 47-90.

5. Rudvin, M., Di Gennaro, E., & Di Rosa, R. T. (2022). Training Language Mediators and Interpreters through Embodied Cognition, Immersive Learning and Virtual Reality: Didactic, Organizational and Cost Benefits. Introduction to the Special Section, Italian Journal of Sociology of Education, 14(3), 131-152. DOI: 10.14658/pupj-ijse-2022-3-6.

6. Lin, T. J., & Lan, Y. J. (2015). Language learning in virtual reality environments: Past, present, and future. Journal of Educational Technology & Society, 18(4), 486-497.

7. Parmaxi, A. (2020). Virtual reality in language learning: a systematic review and implications for research and practice. Cyprus University of Technology. DOI: 10.1080/10494820.2020.1765392.

8. Zhao, G., Fan, M., Yuan, Y., Zhao, F., & Huang, H. (2021). The comparison of teaching efficiency between virtual reality and traditional education in medical education: a systematic review and meta-analysis. Annals of translational medicine, 9(3). 252.

   https://atm.amegroups.com/article/view/61183/html.

Annotation of Qualitative Linguistic Features of Text Readability in Italian Institutional Texts

Nicola Cirillo

University of Salerno
Fisciano, Salerno, Italy
n.cirillo@unisa.it

Alessandro Maisto

University of Salerno

Fisciano, Salerno, Italy
a.maisto@unisa.it

Daniela Vellutino

University of Salerno

Fisciano, Salerno, Italy
d.vellutino@unisa.it

Abstract

This work focuses on the automatic annotation of linguistic readability features in institutional Italian documents. Specifically, our research is focused on qualitative linguistic features identified by other researchers as markers of complexity.

Italian administrative texts often exhibit an excessive complexity that interferes with citizens' right to access information [1, 2, 3]. Many authors identified the features responsible for low readability and proposed rewriting rules [4, 5, 2, 3, 6]. Moreover, in institutional texts, terminology is necessary to communicate information; therefore, an effective simplification system must recognize the institutional terms necessary for communicating the message of public administration [2]. Among these linguistic features, we will focus our research on:

(1) Complex prepositions

(2) Passive voice

(3) Nominalizations and Light Verbs

(4) Double negation

From the Natural Language Processing (NLP) standpoint, there are limited tools available for assessing the readability of Italian sentences. One of the most notable is Read-It [7]. It leverages the Gulpease index [8] and the *Nuovo vocabolario di base* (New Basic Vocabulary) developed by De Mauro [9]. Additionally, Melillo and Maisto [10] introduced a system tailored for evaluating Children's literature, based on Syntactic Complexity –which includes the number of subordinate clauses, deviations from the standard SVO (Subject-Verb-Object) order, and the excessive use of punctuation– and Lexical Simplicity, which considers the frequency of words from the *Nuovo vocabolario di base* [9, 11]. These tools rely mostly on quantitative metrics and provide little information on which portions of the text hinder readability and why. On the contrary, we aim to develop an annotation tool that identifies readability issues in sentences through Nooj dictionaries and grammars. Then, we will evaluate these linguistic resources on a list of sentences extracted from public notices drafted by Italian municipalities [12]

Keywords

Institutional Italian, Automatic Text Simplification, Text Readability, Nooj

References

1. Lubello, S. (2014). Il linguaggio burocratico. Roma, Carocci.
2. Vellutino, D. (2018). L'italiano istituzionale per la comunicazione pubblica. Bologna, Il Mulino.
3. Cortelazzo, M. A. (2021). Il linguaggio amministrativo. Principi e pratiche di modernizzazione. Roma, Carocci.
4. Fioritto A. (1997). Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche. Bologna: Il Mulino.
5. Piemontese, M. E. (1991). "Scrittura e leggibilità: «Due parole»". In Michele A. Cortelazzo (a cura di) Scrivere nella scuola dell'obbligo, Quaderni del Giscel, La Nuova Italia, Firenze, 151-167.
6. Fiorentino G., Ganfi V. (2024 – *in print*). *Parametri per semplificare l'italiano istituzionale: revisione della letteratura*, in ItalianoLinguaDue, Milano, Italy.
7. Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011, July). READ–IT: Assessing readability of Italian texts with a view to text simplification. In Proceedings of the second workshop on speech and language processing for assistive technologies, 73-83.
8. Lucisano, P. & Piemontese, M. E. (1988). Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana. Scuola e città, 110–124.
9. De Mauro, T., & Chiari, I. (2016). Il Nuovo vocabolario di base della lingua italiana. Internazionale. [28/11/2020]. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.
10. Melillo L., & Maisto A. (2024 in print). Valutazione Automatica della Leggibilità dei Testi. Ieri e oggi: la terminologia e le sfide delle Digital Humanities, XXXI Convegno Ass.I.Term (10 dicembre 2021), Verona, Italia.
11. Elia, A., Maisto, A., Melillo, L., & Pelosi, S. (2021). The Lexical Complexity and Basic Vocabulary of the Italian Language. In Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities: 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers 14, 14-23. Springer International Publishing.
12. Vellutino, D., Cirillo N. (2024 – in print). Corpus «Itaist»: Prime Note Per Lo Sviluppo Di Una Risorsa Linguistica Per Lo Studio Dell'italiano Istituzionale Per Il Diritto Di Accesso Civico, in ItalianoLinguaDue, Milano, Italy.

SYNTACTIC AND SEMANTIC RESOURCES

A Set of Disambiguating Grammars for webnooj

Max Silberztein

Université de Franche-Comté

Besançon, France
max.silberztein@univ-fcomte.fr

Abstract

The new webnooj website (https://webnooj.univ-fcomte.fr) allows anyone to apply predefined and user-made NooJ queries to a text to produce a concordance, and perform various statistical analyses. As of now, users can apply queries to an English and a French text; a corresponding Argentinian Spanish page is under construction.

There are many advantages of having access to NooJ functionalities on the Web: the installation of the NooJ software variants (Windows, macOS and LINUX) can be complicated, and the Web application can be accessed from any PC (Windows, macOS or LINUX PC), or even from a tablet or a smartphone, with the exact same simple interface.

As this web site is intended to be used as a demo by a large audience of students and non-linguist specialists, it is important that the results of the queries do not contain false positives. Indeed, it would be very frustrating for a person who is not a linguist and is not familiar with the NooJ approach to apply a query such as "<DET> <N>" to a text, and get a concordance that contains the sequence "(I think) that I (will come back tomorrow)", because "that" *could* be a determiner and "I" could be a noun (even though not in this case).

For this reason, instead of using a very restrictive and "purist" set of linguistic grammars to disambiguate texts without any risk of introducing any mistake, in any text, I have developed a set of 20+ disambiguation grammars that are not necessarily perfect in all cases, but are well adapted to the texts presented on the web site, and are sufficiently robust to be applied to a large number of standard texts. This set of local grammars is structured in 3 subsets; each subset being applied in cascade.

The first set contains mostly filtered lexical entries to get rid of lexical hypotheses that do not appear in standard texts, the second set contains local grammars that are centered around specific unambiguous wordforms (e.g., "is", "the"); the third set contains more abstract patterns such as "<DET+s> <A> <N+s> <V+3+s>".

I will present these local grammars, evaluate the results in terms of the number of annotations removed, and discuss how to improve the disambiguation process.

Keywords

NooJ, Corpus Linguistics, Local grammars, Automatic disambiguation, Automatic tagging, Finite-State Transducers

References

1. Silberztein, M. (Ed.) (2024). Linguistic Resources for Natural Language Processing. Springer Eds: Cham, Switzerland.
2. Silberztein, M. (2016). Formalizing Natural Languages: the NooJ approach. Wiley Eds: Hoboken NJ, USA.

A Collocation Analysis of the Verbs from Lemma <*LAHIR*> Iin Indonesian Translation of the Holy Quran and its Potential Implications for the Construction of NooJ Dictionaries and Grammars of Indonesian

Prihantoro

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

Abstract

Indonesia is one of the largest muslim countries in which Arabic is a foreign language [1]. Therefore, with a population of 270 million people, 87% identify themselves as muslims, the translation of the Holy Quran plays a crucial role. While most studies discuss the holy Quran from theological and religious points of view (such as [2], [3], [4] , it is also important to study the concept from a linguistic point of view. The corpus for the present study was built from the Indonesian translation (ministry of Religion of Indonesia version) of the Holy Quran, indexed and analysed using Sketch Engine and NooJ [5]. I adopted collocation analysis procedure used in Prihantoro [6], [7]. I focused on Indonesian verbs whose lemma is <lahir> 'to give birth' as no previous study about the topic was observed. Concordance lines containing verbs whose lemma is <lahir> were extracted. Several important findings can be highlighted. First, there are two senses found: to give birth (physically) and to explicate something, with the first being the majority. In terms of collocation strength (MI score), collocates relating to the second meaning are stronger. Second, the active form *lahirkan* 'to give birth' is always intransitive, while another active form *melahirkan* 'to give birth' is ambitransitive (with transitive being dominant). One possibility is that Alwi et al.'s [8] argument about the optionality of active voice morpheme meN- (with me- being one of its allomorphs) is optional, and needs to be expanded. We can see that the absence of meN- can mark intransitive verbs. Another possibility is the influence of the corpus being studied. Thus, the pattern is specific only to the Indonesian translation of the Quranic language.

Keywords

Quran, lexicogrammar, Indonesian, Arabic, NooJ, corpus linguistics

References

[1] A. Muradi, F. Mubarak, F. Permana, Y. Hidayat, and I. Wekke, "Revitalization of the Existence of Arabic in Indonesia," SSRN Electron. J., 2021, doi: 10.2139/ssrn.3767448.

[2] A. Aboamro and H. Rizapoor, "Unveiling the Divine Text: Exploring the Analytical Interpretation of the Holy Quran," Integr. J. Res. Arts Humanit., vol. 3, no. 3, pp. 39–48, Jun. 2023, doi: 10.55544/ijrah.3.3.8.

[3] N. M. Almas and A. M. Salh, "Subhani's Interest in Lexical Meanings and Etymology of Quran Translation/ Interpretation (1951- 1990)," J. Univ. Raparin, vol. 9, no. 5, pp. 414–432, Dec. 2022, doi: 10.26750/ Vol(9).No(5).Paper19.

[4] H. Salsabila, F. Muhammad, E. Zulaiha, and M. Y. Firdaus, "Eksplorasi Tafsir Ilmi," Reslaj Relig. Educ. Soc. Laa Roiba J., vol. 5, no. 6, pp. 2797–2807, Jan. 2023, doi: 10.47467/reslaj.v5i6.2595.

[5] M. Silberztein, Formalizing Natural Languages: The NooJ Approach, 1st ed. Wiley, 2016. doi: 10.1002/9781119264125.

[6] P. Prihantoro, "AUTOMATIC RETRIEVAL AND THE FORMALIZATION OF MULTI WORDS EXPRESSIONS WITH F-WORDS IN THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH," J. Hum., vol. 27, no. 2, p. 156, Jan. 2016, doi: 10.22146/jh.v27i2.8709.

[7] P. Prihantoro, "SEMANTIC PROSODY OF WORDS OF EFFECTS IN INDONESIAN," Indones. J. Appl. Linguist., vol. 5, no. 1, p. 101, Jul. 2015, doi: 10.17509/ijal.v5i1.837.

[8]. H. Alwi, H. Lapoliwa, and A. Moelyono, Tata Bahasa Baku Bahasa Indonesia. Badan Pengembangan dan Pembinaan Bahasa, 1998.

Automatic Processing of Free Word Combinations Containing Quantitative Expressions in NooJ

Yauheniya Zianouka

United Institute of Informatics Problems
Minsk, Belarus
evgeniakacan@gmail.com

David Latyshevich

United Institute of Informatics Problems
Minsk, Belarus
david.latyshevich@gmail.com

Yuras Hetsevich

United Institute of Informatics Problems
Minsk, Belarus
yuras.hetsevich@gmail.com

Abstract

The core tasks of speech synthesis and recognition involve a set of operations for text processing, including text normalization and prosodic segmentation. Previous research has focused on syntagmatic delimitation of Belarusian literary texts using NooJ syntactic grammars [1–3]. Their realization is based on the use of rules-based approaches and grammatical patterns for analysing language properties by means of the NooJ environment [4].

Another important aspect is the automatic processing of word combinations containing quantitative expressions (QEMUs) [5]. The goal of normalization is to standardize the representation of these elements (a numeral quantifier – a number or a numeral – and a symbol or word that describes it) to facilitate their annotation and analysis. The study of such units is complicated by the peculiarities of the Belarusian language as an inflectional language [6]. For example, Belarusian has diverse grammatical categories with different morphological forms. In our case, an adjective/numeral must agree with a noun in six cases, number, and gender. It also can be difficult to identify dependencies between words and phrases, as these dependencies are often expressed through morphological forms and inflectional endings.

To accomplish this research, we will create a corpus of scientific-journalistic texts for audio guides. The next step is to extract all quantitative expressions (years, dates, periods, Latin and Roman numerals, percents, and area measurements) using NooJ and create a set of syntactic grammars for expanding them into verbal implementation and separate syntagma. These resources will be integrated into Belarusian NLP applications for various areas, such as syntactic parsing, prosody delimitation, text-to-speech synthesis, and more, serving as a means of voicing audio guides.

Keywords

Prosodic segmentation, Quantitative expression, a Corpus, Automatic processing.

References

1. Hetsevich, Y., Okrut, T. & Lobanov, B. (2016). Grammars for the Sentence into Phrase Segmentation: Punctuation Level. In: T. Okrut, Y. Hetsevich, M. Silberztein, H. Stanislavenka (eds.). International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015": Abstracts (11–13 June, 2015, Minsk, Belarus), Revised selected papers, 74-82. Springer International Publishing.
2. Zianouka Ya., Hetsevich Y., Latyshevich D. & Dzenisiuk Z. (2022). Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module. In: Magali Bigey, Annabel Richeton, Max Silberztein, Izabella Thomas (eds.). 15th International Conference, NooJ 2021, Besançon, France, June 9-11, 2021, Revised Selected Papers, 231-242. Besançon: Springer, Cham.

3. Zianouka, Ya, Hetsevich Yu., Suprunchuk M. & Latyshevich D. (2022). Prosodic Segmentation of Belarusian Texts in NooJ. In: Mariana González, Silvia Susana Reyes, Andrea Rodrigo, Max Silberztein (eds.). 16th International Conference, NooJ 2022, 50-62. Universidad Nacional de Rosari, Rosario : Springer Cham.
4. A Linguistic Development Environment. http://www.nooj4nlp.org/. Last accessed 2022/07/21.
5. Borodina, Ju. & Hetsevich, Yu. (2016). Using NooJ to Process Satellite Data. In: ed. T. Okrut, Y. Hetsevich, M. Silberztein, H. Stanislavenka (eds.). Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015, Minsk, Belarus, June 11-13, 2015, Revised Selected Papers, pp. 182-190. Springer International Publishing.
6. Skopinova, A., Hetsevich, Y. & Borodina, J. (2015). Converting quantitative expressions with measurement unit into an orthographic form and convenient monitoring for Belarusian. In: Monti, J., Silberztein, M., Monteleone, M., Di Buono, M.P. (eds.) Formalising Natural Languages with Nooj 2014, pp. 175-186. Cambridge Scholars Publishing, UK, Newcastle.

Conditional Clauses versus Concessive Clauses in Spanish: An Automatic Treatment with the NooJ Platform

Andrea Fernanda Rodrigo

CETEHIPL
Facultad de Humanidades y Artes, UNR

Rosario, Argentina
andreafrodrigo@yahoo.com.ar

Silvia Susana Reyes

CETEHIPL
Facultad de Humanidades y Artes, UNR

Rosario, Argentina
sisureyes@gmail.com

Mariana González

CETEHIPL
Facultad de Humanidades y Artes, UNR

Rosario, Argentina
marianagonzalez826@gmail.com

Abstract

The CETEHIPL focuses on the pedagogical application of the NooJ tool created by Silberztein [4]. The central ideas of this proposal are developed in our book Aprendo con NooJ [3]. Always based on our Spanish Module Argentina available on the NooJ platform [5], we have been focused on the needs of Spanish language learners whether native speakers or not, and have concentrated on improving their proficiency in the comprehension and production of texts. To achieve this aim it is necessary to review teaching and learning strategies and to integrate information technologies into the automatic processing of subordinate clauses. In this paper we will account for the opposition between two types of syntactic constructions in Spanish (mainly northwestern Argentinian Spanish according to the chosen corpus), concessive [OR_CONCES] and conditional clauses [OR_CONDIC].

In the Nueva Gramática de la Lengua Española de la RAE [2] those structures are identified as pseudo-adverbial, since rather than modifying the verb of the main sentence they actually have implications at a discursive level. The corpus is closely related to our Argentinian culture since it is made up of the lyrics of the folk songs of Atahualpa Yupanqui [6]. In a first descriptive stage, we searched the corpus to check what connectors are used in these subordinate clauses. The conditional connector par excellence is *si* (if) and the most frequently used concessive connector is *aunque* (although). In a second productive stage, it is intended that NooJ recognizes and generates different structures. To this end, after conveniently adapting our dictionaries to process the expressions of Gaucho Spanish, and introducing a new tag [+GAUCH], we will create grammars (for conditional and concessive clauses), seeking to bring together the variety in both prototypical formats, and with a central emphasis on the notion of nucleus phrase proposed by Gabriel Bès [1].

Keywords

Concessive clauses, Conditional clauses, Northern Argentinean Spanish

References

1. Bès, G. G. (1999). La phrase verbal noyau en français. In: Recherches sur le français parlé, vol. 15, pp 273–358. Université de Provence.
2. Nueva Gramática de la Real Academia Española, Espasa Libros, S.L.U., Madrid (2009)
3. Rodrigo, A., Bonino, R. (2019). Aprendo con NooJ: de la Lingüística Computacional a la Enseñanza de la lengua. Editorial Ciudad Gótica, Rosario.
4. Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach. Iste Ediciones. London.
5. Spanish Module Argentina: https://nooj.univ-fcomte.fr/resources.html [last consultation on May 1, 2024]
6. Yupanqui, Atahualpa, available at https://www.letras.com/atahualpa-yupanqui/ [last consultation on May 1, 2024]

Exploring metaliterary cues using NooJ: the case of Les Amours de Psyché et de Cupidon by Jean de La Fontaine

Fabio Meroni

University of Bergamo

Bergamo, Italy
f.meroni2@studenti.unibg.it

Abstract

In contemporary digital humanities, the digitization of literary works involves more than just a change in format; it requires forms of encoding that enable the application of innovative methodologies for studying texts. "Distant reading", as defined by Moretti [6], aims to reveal hidden patterns and prompt new questions and interpretations using existing texts, as this approach complements traditional "close reading" methods [1].

A notable example of this approach in French literature studies is demonstrated by Démoris, Boissieras, and Bermann [2] in their analysis of Marivaux's La Vie de Marianne. They uncovered Marivaux's stylistic use of interrogative clauses and personal pronouns to engage readers in a metafictional game. Similarly, Jean de La Fontaine's Les Amours de Psyché et de Cupidon employs metaliterary devices, yet La Fontaine's lesser-known works lack dedicated scholarly attention and computational linguistic analysis [3].

Building on prior work employing Python [5], this paper explores new avenues using NooJ, leveraging its linguistic resources to enhance empirical analyses. NooJ has proven viable in literary studies contrary to earlier scepticism, Müller [7], Pignot and Piton [8] and Machonis [4] have demonstrated its utility in studying lexical choices and linguistic features in literature.

In the same way, the aim of this study will be the identification of those linguistic resources that La Fontaine employs to deepen the focus on the metafictional purpose of his work. Starting from the disambiguation of the lexical entries "Amour" and "amour" with specific vocabulary-based constraints, specific queries will be formulated to examine instances where La Fontaine interacts with the reader. XML tagging will distinguish the narrative frame from the main narration, facilitating focused analysis of La Fontaine's techniques for breaking the fourth wall.

Keywords

Digital Humanities, Distant Reading, Literary Analysis, Narrative Content Analysis, Named Entity Recognition

References

1. Bode, K. (2017). The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. *Modern Language Quarterly*, 78(1), 77–106.
2. Démoris, R., Boissieras, F., Bermann, M. (2014). *Marivaux, La Vie de Marianne*. Atlande.
3. La Fontaine, J. (1669). *Les Amours de Psyché et de Cupidon*. Chez Claude Barbin.

4. Machonis, P.A. (2021). Where the Dickens Are Melville's Phrasal Verbs?. In B.Bekavac, K. Kocijan, M. Silberztein, & K. Šojat (Eds.), *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2020* (1st ed., pp. 203–213). Springer.

5. Meroni, F. (2024). [Unpublished manuscript] La vocation métalittéraire de l'amour et de la curiosité dans *Les Amours de Psyché et de Cupidon* de Jean de La Fontaine. To feature in *InScriptum*, 5.

6. Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, 1(2), 54–68.

7. Müller, R. (2013). Nooj as Concordancer in Computer-Assisted Textual Analysis: The Case of the German Module. In S. Koeva, S. Mesfar, M. Silberztein (Eds.), *Formalising Natural Languages with NooJ 2013: Selected papers from the NooJ 2013 International Conference* (1st ed., pp. 203–213). Cambridge Scholars Publishing.

8. Pignot, H., Piton, O. (2014). Mary Astell's Words in *A Serious Proposal to the Ladies* (part I), A Lexicographic Inquiry with NooJ. *ArXiv*, https://arxiv.org/abs/1412.1215.

Who's done it? He or she crime?

Filip Lukeš

Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
filip.lukes@hotmail.com

Filip Vuković

Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
filip.vukovic99@gmail.com

Anela Rački

Faculty of Humanities and Social Sciences
University of Zagreb
Zagreb, Croatia
rackianela9@gmail.com

Abstract

This project deals with the differences in the vocabulary used to describe crimes with regard to the gender of the perpetrator. The analysis was carried out using online media articles produced between 2011 and 2023. The vocabulary used to describe the crime was analyzed under the assumption that the description is more harsh when the perpetrator is a man, regardless of the severity of the crime. The project seeks to reveal the character used to describe the crime as well as the perpetrator themself. Three categories were introduced to describe the nature of the article: neutral, negative and harsh. Regardless of the type of crime committed, it is assumed that the nature of the text is predominantly neutral, while negative and harsh texts presuppose the gender of the perpetrator. Our starting hypothesis is that most articles are written neutrally. Our second hypothesis is that, if not written neutrally, articles are harsher towards male perpetrators.

Keywords

Vocabulary varieties, NooJ vocabulary analysis, Crime description difference

References

1. Gueta, K., & Chen, G. (2016). Men and women inmates' accounts of their pathways to crime: A gender analysis. Deviant behavior, 37(12), 1459-1472.
2. Nawaz, T., & Hussain, I. (2021). A forensic linguistic analysis of language of news crime reports in Pakistan. Harf-o-Sukhan, 5(4), 585-595.
3. Soro, S. H. (2019). Analysis of English Vocabulary in the Mind of Students. International Journal of Nusantara Islam, 7(2), 220-229.
4. Moody, S., Hu, X., Kuo, L. J., Jouhar, M., Xu, Z., & Lee, S. (2018). Vocabulary instruction: A critical analysis of theories, research, and practice. Education Sciences, 8(4), 180.
5. Adams, M. (2013). Vocabulary analysis in sociolinguistic research. Research methods in sociolinguistics: A practical guide, 163-176.
6. Deane, P. (2012). NLP methods for supporting vocabulary analysis. Reaching an understanding: Innovations in how we view reading assessment, 117-144.
7. Allen, L. K., & McNamara, D. S. (2015). You Are Your Words: Modeling Students' Vocabulary Knowledge with Natural Language Processing Tools. International Educational Data Mining Society.

Construction of a Translation Relative Clauses System with NooJ

Nizar Jarray
ISG Gabes
University of Gabes
Gabes, Tunisia
nizar.jarray1998@gmail.com

Héla Fehri
MIRACL Laboratory
University of Sfax
Sfax, Tunisia
hela.fehri@yahoo.fr

Hajer Cheikhrouhou
Language and Treatment
Automatic Laboratory
University of Sfax
Stax, Tunisia
hajer.cheikhrouhou369@gmail.com

Abstract

As scientific research continues to expand at a rapid pace, there arises an ever-increasing need for sophisticated automatic translation approaches. In fact, translation plays a vital role in facilitating communication across languages and cultures, enabling researchers to access and disseminate knowledge more effectively. Automatic translation faces significant challenges in accurately handling relative phrases, which establish connections between different sentence elements and can vary greatly across languages. Translating these phrases requires a deep understanding of the grammatical structures and contextual nuances of both source and target languages. Furthermore, the complexity of relative clauses, often involving ambiguous references and intricate syntax, adds another layer of difficulty to the translation process. Consequently, automatic translation systems must navigate these complexities to maintain coherence and fluency in the translated text. Addressing these challenges is crucial for enhancing the accuracy and reliability of automatic translation, thereby improving its usability and effectiveness in facilitating communication across different languages. This paper aims to propose a method for translating French relative phrases into Arabic. By leveraging the linguistic analysis capabilities of the NooJ platform, we propose a method to address these challenges and achieve accurate translations from French to Arabic. Our method involves analyzing the grammatical structures and contextual nuances of both languages, identifying and processing relative clauses, and applying appropriate translation rules within the NooJ environment. Our proposal is based on transducers and dictionaries. Through this method, we aim to enhance the efficiency and effectiveness of automatic translation systems in handling relative phrases, thereby facilitating cross-linguistic communication between French and Arabic speakers.

Keywords

Translation, NooJ, Transducer, Relatives Clauses, Arabic, French

References

1. Cheikhrouhou, H. (2017, June): Le verbe et la problématique de la traduction automatique français-arabe (utilisation de la plate-forme NooJ).
2. Lafhej I. (2007). Aspects de variations entre le français et l'arabe : Le cas des subordonnées relatives. Revue Maghrébine des Langues, 5(1), 238-255.

NLP SOFTWARE APPLICATIONS

Linguistic Analysis for Legal Natural Language Question Analysis in Question Answering System Development

Ismahane Kourtin

CRIT
Université de Franche-Comté

Besançon, France
kourtin_ismahane.math@yahoo.fr

Samir Mbarki

Ibn Tofail University

Kenitra, Morocco
mbarkisamir@hotmail.com

Abstract

The growing need for efficient and accurate legal information retrieval in Morocco, which utilizes both French and Arabic languages, has fueled the development of Question Answering systems (QAS) [1] tailored for the legal domain. This paper focuses on the crucial initial phase of QAS development, specifically addressing the linguistic analysis required for effective Natural Language Question Analysis (NLQA) in the legal context. The objective is to enhance the system's ability to comprehend and respond to complex legal queries posed in natural language in French.

The legal domain poses unique challenges for Natural Language Question Answering (NLQA) systems due to its intricate vocabulary, complex sentence structures, and reliance on specific legal interpretations. Linguistic analysis becomes even more critical in this context, offering a powerful tool to unlock the nuances of legal language and improve question understanding within NLQA systems.

The study delves into the unique challenges posed by legal language, characterized by its precision, ambiguity, and reliance on domain-specific terminology [2]. By employing advanced linguistic analysis techniques, including syntactic, semantic, and pragmatic approaches, the paper aims to overcome these challenges and provide a robust foundation for NLQA. Special attention is given to the identification and disambiguation of legal terms, as well as the recognition of syntactic structures and semantic relationships within legal questions.

To analyze user questions in natural language, we opted for a pattern-based linguistic approach using the NooJ natural language processing platform [3,4]. NooJ allows us to build, test, and manage formal descriptions in a wide coverage of natural languages, in the form of dictionaries and electronic grammars.

Keywords

Question-answering system (QAS), Natural language processing (NLP), NooJ, Legal field

References

1. Hirschman, L. & Gaizauskas, R. (2001). Natural language question answering: the view from here. Natural Language Engineering, 7(4), 275–300. URL http://dl.acm.org/citation.cfm?id=973891.
2. Katz, D. M. & Hartung, D. (2023). Natural Language Processing in the Legal Domain. https://arxiv.org/abs/2302.12039.
3. Silberztein, M. (2003) NooJ Manual.
4. Silberztein, M. (2016). Formalizing Natural Languages: the NooJ approach. Wiley-ISTE Ed.: Hoboken, NJ, USA.

NooJ Digital Text Automation Workshop: Artificial Intelligence (AI): Data Processing Automation and Cybersecurity Technologies

Ritamaria Bucciarelli

Università Cà Foscari

Venice, Italy
rita.bucciarelli@unive.it

Francesco Terrone

SIDELMED (Spa)

Taranto, Italy
francesco.terrone@sidelmed.it

Samuela Franceschini

Università Cà Foscari

Venice, Italy
.franceschini@unive.it

Andrea Fernanda Rodrigo

Universidad Nacional de Rosario

Rosario, Argentina
andreafrodrigo@yahoo.com.ar

Javier Julian Enriquez

Universitat Politècnica València

Valencia, Spain

jajuen@alumni.upv.es

Abstract

The research project, which will be conducted from 2020 to 2024, is being carried out by a team of experts. The project is based on models, processes, methods, and technologies for NLP2023. The research topic for the project is: the field of study known as Natural Language Processing, or NLP, which is concerned with the processing of human language in digital form [1]. This project proposes communication strategies among scientists for cooperative actions in real time. Text automation is the primary focus of this subject, which involves the transition from human to digital thinking. The irrefutable statements about the formal language (i.e. synthetic and fixed, and homology classes), as proposed by Douglas R [2] and by M. Planat, are presented in Graph Coverings for Investigating Non-Local Structures in Proteins, Music, and Poems [3].

The research into the discussion is based on point 0, which converts alphabetical code with relative grammar to formal grammar. The methodology of the process is described by Bucciarelli. et al., during EUPORIA 2021 [4], [5] when quantum physicists are invited to transfer information from quantum theory of phonic musicology to phonic literature.

The actions for automatic processing, analysis, and production of scans in environments and models was made by:

- Team DM - Data surveys and statistical calculations and prediction of linguistic (phono-tonal) and manipulative mechanisms on the text under study according to the required type (literary)
- NooJ M. Silberztein Laboratory - for the development of automatic data analysis, for production of approximation graphs for DB construction [6], [7].
- Quantum computing is used for data automation, as stated by C. Mereghetti, to construct finished automata in …. [8]
- Ettore Galluccio et al. affirm in SQL Injection, Strategies Pack<t> in Experts in secure system design and cyber risk management…. [9].

Keywords

Automation, Numerical groups, Quantum computing automata, Acronyms.

References

1. R. Bucciarelli, Research project 2020-2023: model -processes- methods -technologies NLP-.
2. Douglas, R. G. (1980). C*-Algebra Extensions and K-Homology. (AM-95). Princeton University Press. http://www.jstor.org/stable/j.ctt1b9s045
3. Planat, M., Aschheim, R., Amaral, M. M., Fang, F., & Irwin, K. (2021). Graph Coverings for Investigating Non-Local Structures in Proteins, Music, and Poems. Sci, 3(4), 39. https://www.mdpi.com/2413-4155/3/4/39 .
4. Bucciarelli, R., Veronesi, I., Greco, M., Gagliardi, N., Tortoriello, F. S., Savarese, G., ... & Villari, P. (2021, February). Communication in Congress Webseminar Euporia 2021. In webseminar-Euporia 2021, Venice, Italy. ⟨hal-04065553⟩.
5. Bucciarelli, R., Veronesi, I., Greco, M., Gagliardi, N., Tortoriello, F. S., Savarese, G., ... & Villari, P. (2021, February). Communication in Congress Webseminar Euporia 2021: Homero, PowerPoint Presentation. Webseminar-Euporia 2021, Venice, Italy. ⟨hal-04065553⟩.
6. Silberztein, M. (2016). Formalizing natural languages: The NooJ approach. John Wiley & Sons.
7. Silberztein, M. (2018). A new linguistic engine for NooJ: Parsing context-sensitive grammars with finite-state machines. In Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications: 11th International Conference, NooJ 2017, Kenitra and Rabat, Morocco, May 18–20, 2017, Revised Selected Papers 11 (pp. 240-250). Springer International Publishing.
8. Bertoni, A., Mereghetti, C., & Palano, B. (2003). Quantum computing: 1-way quantum automata. In Developments in Language Theory: 7th International Conference, DLT 2003 Szeged, Hungary, July 7–11, 2003 Proceedings 7 (pp. 1-20). Springer Berlin Heidelberg.
9. https://www.researchgate.net/publication/221212400_Quantum_Computing_1-Way_Quantum_Automata.
10. Galluccio, E., Caselli, E., & Lombari, G. (2020). SQL Injection Strategies: Practical techniques to secure old vulnerabilities against modern attacks. Packt Publishing Ltd.

From the social network data to the document-oriented NoSQL Datawarehouse using NooJ platform

Amira Abbes
MIRACL LABORATORY
Sfax, Tunisia
amyyraa29@gmail.com

Héla Fehri
MIRACL LABORATORY
Sfax, Tunisia
hela.fehri@yahoo.fr

Ines Ben Messaud
MIRACL LABORATORY
Sfax, Tunisia
ben.messaoud.ines@gmail.com

Abstract

In the last decade, social media has significantly changed the world. In fact, the rapid and vast adoption of these platforms has influenced human behavior and society in general. All these media channels are essential in our daily lives because they enable us to easily communicate, interact and share a variety of content, with one another [1]. Nevertheless, social media plays a crucial role across various domains, particularly on opinion analysis area with decisional systems, allowing companies to enhance their products from reviews posted on social media such as Facebook, YouTube, etc... In this respect, there is a growing interest in opinion mining expressed on social media for business decision support [2].

The aim of this paper is to present a method to construct data warehouse from social media, by implementing document-oriented NoSQL databases based on a linguistic approach using resources of the NooJ platform [3]. The proposed method is composed of 4 steps: "Data Extraction", "Data Classification", "Multidimensional Modeling" and "Data Loading". Data extraction is carried out using a tool called FacePager, which facilitates the extraction of comments from a Facebook page. Subsequently, these comments are filtered and categorized into positive, negative, and neutral sentiments using the NooJ platform. Indeed, from the file that contains the comments, we have identified the necessary semantic features that allow us to classify the opinions. We have also identified the patterns that we need in order to classify the comments. These patterns have been transformed into transducers. Our warehouse is finally modeled and loaded by these classified comments.

To assess and validate the methodology, we built a corpus containing over 100 comments related to the technology domain. As resources, we used the "_dm" dictionary with the addition of the semantic features (positive, negative and neutral). Additionally, we used our own dictionary "marque", encompassing various names of both Samsung and Apple, as well as names of other brands (Huawei, Redmi, etc.). We also used three main syntactic grammars that we constructed respectively for the classification of positive, negative and neutral opinions. The obtained results are satisfactory and encouraging.

Keywords

Data warehouse, Social media, Opinion analysis, NoSQL database, Transducer

References

1. Moalla, I., Nabli, A., & Hammami, M. (2022). Data warehouse building to support opinion analysis in social media. Social Network Analysis and Mining, 1-15.
2. Moalla, I. N., Moalla, I., Nabli, A., & Hammami, M. (2018). Towards Opinions analysis method from social media for multidimensional analysis. Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia, 8-14.
3. Silberztein, M. (2015). La formalisation des langues : l'approche NooJ. Collection Sciences Cognitive et Management des Connaissances. Edition ISTE, London

Nooj Grammars and Cybersecurity. Possible Future Uses and Scenarios

Beatrice Folino

Epicode Italia

Rome – Milan – Berlin
beatricefolino@outlook.it

Mario Monteleone

Università degli Studi di Salerno

Fisciano, Salerno, Italy
mmonteleone@unisa.it

Abstract

According to the Italian Observatory of Digital Innovation [1], "Cybersecurity is the main digital investment priority in Italy, also thanks to the push of the PNRR and the guidance of the National Cybersecurity Agency. However, the relationship between spending on cybersecurity and GDP is 0.10%, still far from the other G7 countries: 67% of companies in Italy note an increase in attack attempts, 61% have increased their budget for IT security. In 53% of companies, there is a formalized Chief Information Security Officer. In 80% structured training plans on attack risks. Only a third of companies have methodologies for financial quantification of cyber risk. Cyber attacks are continually increasing, with 1,141 serious incidents detected by Clusit in the first half of 2022 alone, +8.4% compared to the same period in 2021, and threats are increasingly also affecting critical infrastructures. In this context, 67% of companies note an increase in attack attempts and 14% have suffered tangible consequences following cyber incidents, such as service interruptions, delays in process operations or reputational damage. Certainly, in Italy attention is growing for cybersecurity, which in 2023 is confirmed as the main digital investment priority among companies, both large and SMEs. As many as 61% of organizations over 250 employees have decided to increase their budget for cybersecurity activities in the last 12 months. Overall, in 2022, the Italian cybersecurity market will reach a value of 1.86 billion euros, with an exceptional acceleration of +18% compared to 2021. The ratio between spending on cybersecurity and GDP in Italy stands at 0.10%, slightly up compared to 0.08% the previous year. However, this is a result that places our country in last place among those in the G7. The ranking is led by the United States and the United Kingdom, with a ratio of 0.31%. For France and Germany the ratio is 0.19% and 0.18% respectively." At present, Cybersecurity routines mostly use encryption algorithms. According to Rootstrap [2], "An encryption algorithm is a mathematical procedure or set of rules that can either perform or determine how data is to be encrypted. An algorithm converts information into meaningless ciphertext and requires a key to transform the data into its original form. Simply put, the best encryption algorithms are programs that easily allow a sender to scramble their data and then permit a receiver to unscramble it on the other end. […] Many different encryption algorithms rely on various methods of encryption. The following are the best encryption algorithms worth looking into or employing […]: 3DES Algorithm, AES Algorithm, RSA Algorithm, ECC Algorithm, BlowFish and TwoFish. It is important to note that these algorithms primarily rely upon the properties of mathematics."

However, as stated by Nvdia researchers [3], "… the role of Natural Language Processing […] in the cybersecurity-world starts to gain traction. So what may be the role and applications of NLP in cybersecurity? […] people started communicating with machines through constructed language (e.g. programming languages, etc.) but are increasingly using the natural language to do it (e.g. chatbots, virtual assistants, etc.). [Therefore] here are some NLP tasks that may be considered in creating threat or attack tools, or in defending humans and machines:

Language Analytics: language modelling, sentiment analysis, text classification, named-entity recognition, natural language inference, relation extraction, semantic parsing, co-reference resolution, entity linking, relational reasoning, semantic composition, language identification and translation, entity and information extraction, intent detection and

classification, stance and fake news detection, rumor detection, hate speech detection, clickbait detection, abuse detection.

Language Generators: question-answering systems, text and dialogues generation, text summarization, slot filling for knowledge base population tasks, scripts and programming code generation.

Hence, considering the Cybersecurity NLP needs highlighted by Nvidia Researchers, we will verify in this study if and how it is possible to insert (parts of) Cybersecurity algorithms codes into specific NooJ grammars, in order to use NooJ as a device to detect, highlight and challenge cyber-attacks automatically and in real time. Ultimately, we will also try to outline the structure of a "standardized" automatic NLP system to support the management of similar attacks.

Keywords

NooJ, NooJ Grammars, NooJ Finite-State Automata and Transducers, Cybersecurity Algorithms, Unix Network Security Model, Apache Web Server Security

References

1. Web page, https://www.osservatori.net/it/ricerche/comunicati-stampa/cybersecurity-italia-mercato-crescita. Last accessed: February 8, 2024.
2. Web Page, https://www.rootstrap.com/blog/the-best-encryption-algorithms-for-cyber-security. Last accessed: February 13, 2024.
3. Web page, https://developer.nvidia.com/blog/changing-cybersecurity-with-natural-language-processing/. Last accessed: February 8, 2024.
4. Silberztein, M.: The NooJ Manual. NooJ Association Web Page, http://www.nooj-association.org last accessed 2023/19/03.
5. Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach. ISTE Ltd and Wiley, London.

Beneath Sentiment Analysis: A Comparison of Polarity Scoring Approaches

Luigi Celardo

University of Naples Federico II
Naples, Italy
luigi.celardo@unina.it

Giuseppe Giordano

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

Michelangelo Misuraca

University of Calabria
Arcavacata di Rende
michelangelo.misuraca@unical.it

Mario Monteleone

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

Abstract

The technological progresses, as well as the increasing diffusion of Web 2.0 during the last few years, transformed the way people use to communicate feelings, opinions and experiences about the diverse aspects of everyday-life. More and more, social media, online forums and review sites are used to share subjective standpoints and exchange information. Since comments and reviews are written as free texts, information concerning the opinions are encoded in a form difficult to process automatically. In a Natural Language Processing (NLP) perspective, it is possible to pre-treat the collected textual body to transform the unstructured data into structured data, and subsequently perform statistical analyses to extract and manage the underlying knowledge base. Dealing with opinions, it is particularly interesting to consider the semantic orientation of texts, expressing in a numerical form the so-called sentiment. Several alternative approaches have been proposed concerning how to determine the negative/positive orientation embodied in the texts and use these scores in an opinion mining strategy. Once the collection of texts has been tagged with the corresponding sentiment, it is possible to perform a supervised classification procedure to automatically categorise new cases. In this work we aim at comparing three of these different alternatives. Firstly, we consider a lexicon-based polarity scoring, employing a customised lexicon of Italian lemmas derived by known resources developed, for example, in the Sentix project [1], in the Opener project [5] and other studies [3]. Secondly, we consider a linguistic model trained on FEEL-IT [2], an Italian reference corpus used to develop the UmBERTo model [4]. This kind of model (based on the so-called transformer encoding) uses neural networks with several hidden layers (i.e., with a deep learning perspective) to analyse and process texts. Its main advantage is the possibility to produce a different embedding for each word depending on the given context, whereas other approaches are not able to take into account this kind of information. Thirdly, we use the NooJ environment [6] [7] to attribute the sentiment by means of specific NooJ Local grammars accounting for the lexical/morphological, syntactic, and semantic features necessary to perform and achieve the analysis. The different solutions for polarity scoring ad sentiment detection are tested on a collection of reviews on tourist point of interests in the city of Naples (Italy) scraped from TripAdvisor, since digital platforms hosting reviews of tourist attractions and Point of Interests (PoI) are becoming increasingly significant both in the industrial and the scientific domain

Keywords

NLP, NooJ, NooJ sentiment analysis, Pre-trained language, Text transformers

References

1. Basile, V., Nissim, M. (2013). Sentiment analysis on Italian tweets. In A. Balahur, E. van der Goot, A. Montoyo (eds.), Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 100-107. Association for Computational Linguistics.
2. Bianchi F., Nozza, D., Hovy, D. (2021). FEEL-IT: Emotion and Sentiment Classification for the Italian Language. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 76–83. Association for Computational Linguistics.
3. Bolasco, S., Della Ratta, F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. In G. Purnelle, C. Fairon, A. Dister (eds.), Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles, 202-210. UCL Presses Universitaires de Louvain.
4. Devlin J., Chang M-W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2 [cs.CL] 24 May 2019.
5. Russo, I., Frontini, F., Quochi, V. (2016). OpeNER Sentiment Lexicon Italian - LMF, Institute for Computational Linguistics "A. Zampolli", National Research Council, Pisa.
6. Silberztein, M. (2003). The NooJ Manual. NooJ Association Web Page, http://www.nooj-association.org last accessed 2023/19/03.
7. Silberztein, M., Formalizing Natural Languages. (2016). The NooJ Approach. ISTE Ltd and Wiley, London.

CORPUS LINGUISTICS & DISCOURSE ANALYSIS

Using NooJ in action research: the case of personal expression in professional situations

Magali Bigey

ELLIADD, University of Franche-Comté, *Mediation and Digital Practices* research team

Besancon, france

magali.bigey@gmail.com

Abstract

The action research we propose to present explores the dimensions of well-being at work, going beyond the absence of ill-being to include satisfaction, fulfillment and commitment. Against a backdrop of rapid transformation in the world of work, the study focuses on issues specific to well-being in the workplace, highlighting the complexities and actions required. The focus of this research is on the merger of two companies in the water sector of the city of Lyon, generating tensions and questioning. To analyze this situation, the ValUm project, carried out in partnership between the University of Franche-Comté and Umanao Experience, uses NooJ to analyze work situations collected from the company's 310 employees, during a day when they were brought together in the same place. This linguistic approach adds an extra dimension to the analysis, enabling in-depth formalization of the linguistic data collected as part of the research. The data collected focuses on employees' individual perceptions of their jobs and the company, highlighting that well-being at work goes beyond material satisfaction. It encompasses aspects such as personal fulfillment, psychological health, engagement, and work-life balance, all of which are linked to personal work-related values. This multidimensional approach requires a holistic understanding, facilitated by the use of NooJ to process large-scale texts and corpora. The increasing digitization of professional activities is also taken into account in the analysis. The text highlights the importance of striking a balance between the advantages of digital technologies in terms of efficiency and flexibility, and challenges such as information overload, constant pressure, and the dehumanization of interactions. NooJ plays a crucial role in this approach, enabling a fine-grained analysis of linguistic data linked to the impact of digital technologies on workplace well-being and values.

The ValUm project, supported by the use of NooJ, aims to identify unifying values, supportive policies and management practices to create fulfilling work environments despite profound changes in workspaces. This proposal also highlights the importance of interpersonal communication and the fluidity of exchanges in the quest for humane and sustainable workplaces.

Keywords

Well-being at work, Workplace satisfaction, Employee engagement, Digital technology impact

References

1. Bertrand F. & Hansez I., (2010). Facteurs d'insatisfaction incitant au départ et intention de quitter le travail : analyse comparative des groupes d'âges, Le travail humain, 73(3), 213-237.
2. Guillemin, M., (2018). Le bonheur au travail : qu'en disent les scientifiques ?, Environnement, Risques & Santé, 17, 487-497.
3. Hernandez, J. R., Lhuilier, D., Araujo, N. G., & Pujol, A. (2021). Mal-être et bien-être au travail. Introduction. Bulletin de psychologie, 74(2), 67-69.
4. Silberztein, M. (2015). La formalisation des langues : l'approche de NooJ. ISTE: Londres.

"Are you the man of culture?" Analysis of YouTube Comments on Male and Female Athletes

Bruno Križić

Faculty of Humanities and Social Sciences

Zagreb, Croatia
brunokrizic@gmail.com

Patricija Petanjek

Faculty of Humanities and Social Sciences

Zagreb, Croatia
pa3cija.petanjek@gmail.com

Abstract

The seemingly lighthearted term "man of culture," used in YouTube comments, can mask a concerning trend: the sexualization of female athletes. This study investigates gendered representation in comments on male and female athletes across football, swimming, and athletics. These sports were chosen for their popularity, diverse participation across genders, and contrasting gender perceptions within sports culture. As Kane [1] argues, "Language both reflects and shapes our understandings of gender in sport." Comments praising female athletes for their appearance rather than their athletic prowess exemplify this dynamic.

Building on Messner's [2] concept of "hegemonic masculinity" in sports, we hypothesized a higher frequency of sexist language directed towards female athletes. Using NooJ, a linguistic analysis tool, we analyzed a dataset of 22,000 comments. Our findings confirmed this expectation, with the most sexist comments targeting female athletes in athletics. This aligns with the potential for such sports to objectify female physiques. Interestingly, the most positive comments were for female swimmers, perhaps reflecting the focus on technical skill in that sport, where aesthetics play a lesser role. Conversely, the most negative comments targeted male football players, suggesting a different dynamic of online criticism, possibly centered on performance or team loyalty.

This research highlights the nuanced ways language portrays male and female athletes. It underscores the importance of fostering balanced and gender-neutral discussions of athletes online. Moving forward, efforts are needed to promote respectful online sports commentary that celebrates athletic achievement for all genders. Social media platforms also have a responsibility to create environments that discourage sexist language and promote inclusivity.

Keywords

Youtube comments, Gender, Sexism, Sports, Online discourse, Gendered language

References

1. Kane, M. J. (2006). Language and gender in sport. Routledge.
2. Messner, M. A. (2002). Men're still standing: The masculinity of sports. Illinois: University of Illinois Press.

Drama of the Heart: A Journey into Emotional Dynamics in Anna Karenina

Lucija Bažant

Faculty of Humanities and Social Sciences

Zagreb, Croatia
ibazant@m.ffzg.hr

Mate Grizelj

Faculty of Humanities and Social Sciences

Zagreb, Croatia
matgrize@m.ffzg.hr

Abstract

Emotion articulation is a complex and context-sensitive process. There are several reasons for this [1], including the complexity of emotions themselves. Often, people want to express more than one emotion at once, or they may not know exactly how they feel. Additionally, the linguistic use of emotional terms can be confusing, adding to the complexity of emotion expression. There is much work to be done in emotion recognition, which is why we have chosen to analyse the mechanisms behind emotion expression. By doing so, we hope to gain a better understanding of emotions and various relationships in this novel, as well as in Tolstoy's other works.

This paper analyses the mechanisms through which emotions are articulated by Tolstoy's characters in his novel Anna Karenina. The research is conducted using the Croatian translation of the novel by Martin Lovrenčević, and employs the NLP tool NooJ for annotating words indicative of emotions, classifying them as positive or negative. Given resource constraints, our focus lies on identifying verbs related to emotions, particularly psychological verbs sourced from the e-Glava database of Croatian verb valences. To conduct this research, we initially acquired the complete translated text of Anna Karenina and proceeded to segment the text into chapters. The primary language of the text is Croatian, but French and English are also included due to the nature of the novel. We conducted our analysis only on Croatian words.

Given the thematic context of Anna Karenina, our anticipation is a prevalence of positive emotions, particularly those intertwined with the theme of love. The novel exhibits a greater frequency of positive emotions over negative ones, although this observation lacks statistical significance. A statistically significant prevalence of positive emotions is evident solely in the first chapter, aligning with the thematic essence of the novel.

This research project draws inspiration from analogous study that has explored happiness, i.e. the mechanisms through which happiness is articulated.

Keywords

Emotion articulation, Anna Karenina, Psychological verbs, Croatian language

References

1. Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). Emotion detection in text: a review. arXiv preprint arXiv: 1806.00674.

Climate Change and Migration: Government Officials' Narratives at COP26 and COP28

Jiwon Shin

X23 Science in Society

Treviglio, Italy
jiwon.shin@x-23.org

Abstract

Climate change-induced displacement and migration is a type of movement that arises from slow and rapid-onset events regarding climate change [1]. In the era of the climate crisis, the phenomena encompass drought, floods, elevated temperatures, and a growing number of unusual weather occurrences. The repercussions of climate change are primarily noticeable in the Global South, which is the main area of analysis, as a result of the power disparity between the Global North and the Global South [2]. This article attempts to present a comparative study of contemporary narratives encapsulated in the outcomes of COP26 and COP28, the United Nations Climate Change Conference (COP) in 2021 and 2023, respectively. Dissecting the core concepts, tones, and techniques employed to describe the environmentally induced migration, the author endeavours to shed light on the proceedings of COP as a site for the social construction of narratives and elucidates the international political actions initiated in response to the climate challenge. The main objective of this article is to examine speeches delivered by government representatives from the Global South at the conference. Adopting NooJ, a robust Natural Language Processing (NLP) tool for discourse analysis within a social constructivism framework, can offer insights into uncovering linkages in speeches and narratives formed by government officials through a multi-level analysis of linguistic phenomena [3]. Probing the gist of the argument, it explores the intricate nexus between migration and the climate crisis, as well as the political narratives woven into the fabric of COP26 and COP28.

Keywords

Climate change-induced displacement, Climate migration, Climate Change, COP, NooJ

References

1. Kaczan, D. J., & Orgill-Meyer, J. (2020). The impact of climate change on migration: a synthesis of recent empirical insights. Climatic Change, 158(3), 281-300.
2. Faist, T. (2020). Commentary: Getting out of the climate migration ghetto: Understanding climate degradation and migration as processes of social inequalities. International Migration, 58(6), 247-253. https://doi.org/10.1111/imig.12793
3. Silberztein, M. (2016) Formalizing Natural Languages: The NooJ Approach. Wiley-ISTE, London.

Corpus-Based Exploration of Defense-Related Political Speeches

Maja Ivanović

Faculty of Humanities and Social Sciences University of Zagreb

Zagreb, Croatia
maivanovi@m.ffzg.hr

Kristina Kocijan

Faculty of Humanities and Social Sciences University of Zagreb

Zagreb, Croatia
krkocijan@ffzg.unizg.hr

Abstract

Analyzing the relevant domain of political discourse is not an easy feat as its subject matter is both extensive and ambiguous [3]. Furthermore, a wide array of analytic techniques often relying on qualitative research methods and methodology presents another challenge. Thus, Natural Language Processing (NLP) tools have gained prominence, not only as a means of supplementing qualitative analysis by introducing quantitative support [7], but also specifically when analyzing political documents, texts [6], and speeches held by various politicians, including international leaders and heads of state. The purpose of this work is to lay the foundation for the preparatory phase of a larger study, the first step of which includes the familiarization with the basic concepts in speeches related to the defense realm held by Chinese President Xi Jinping from 2012 to 2019, and which were officially translated into English and collected in three volumes of his speeches called The Governance of China I, II and III.

The aim is to chronologically screen the speeches to classify the words and terms related to (national) defense, combat preparedness, and war in order to detect marker words by utilizing open-source packages; NooJ: a corpus processor and linguistic engine for developing NLP software applications, and Voyant Tools: a web-based reading and analysis environment for digital texts [1]. Additionally, we will utilize the categorization of international events from the research conducted by Goldstein and Schrodt *et al*. as a reference point in the initial processing steps of this study [2], [4], [5] and [8]. We believe that this paper could be used as the first phase for the future exploration of political speeches with open-source packages, such as NooJ and Voyant Tools. Moreover, we will show that NooJ and Voyant Tools could be utilized jointly as complementary digital environments suitable for exploring political discourse.

Keywords

Content analysis, Marker words, Political speeches, NooJ, Voyant tools

References

1. Bednarek, M., & Carr, G. (2021). Computer-Assisted Digital Text Analysis for Journalism and Communications Research. Introducing Corpus Linguistic Techniques That Do Not Require Programming. Media International Australia, 181(1), 131-151.
2. Goldstein, J. S. (1992). A Conflict-Cooperation Scale for WEIS Events Data. *The Journal of Conflict Resolution*, 36 (2). http://www.jstor.org/stable/174480
3. Chilton, P. (2003). *Analyzing Political Discourse. Theory and Practice*. London. Routledge. 1st Edition.
4. Halterman, A., Bagozzi, B. E., Beger, A., Schrodt, P., Scraborough, G. (2023). PLOVER and POLECAT: A New Political Event Ontology and Dataset. In SocArXiv. https://osf.io/preprints/socarxiv/rm5dw/download

5. Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global Data on Events, Location, and Tone. 1979–2012. ISA Annual Convention. Volume 2. No. 4.

6. Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2), 254–277.

7. Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. London. ISTE Ltd. Hoboken. John Wiley & Sons, Inc. 1st Edition.

8. Schrodt, P. A., Yilmaz, O., Gerner, D. J., & Hermreck, D. (2008). Coding Sub-State Actors Using the CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework. Paper presented at the Annual Meeting of the International Studies Association, San Francisco, CA.

How (not) to Talk about Femicide and Gender-based Violence: An Analysis of Narrative Strategies Adopted by Italian Newspapers

Lisa Pezzali

University of Bergamo Bergamo, Italy
l.pezzali1@studenti.unibg.it

Abstract

On an annual basis, starting from 2015, the renowned and esteemed *Istituto dell'Enciclopedia Italiana Treccani* (Institute of the Italian Encyclopaedia Treccani) has been organising a transmedia campaign, wherein the 'word of the year' is chosen. The primary objective of the initiative lies in promoting a correct and conscious use of the Italian language and its lexical components. In 2023, the selected term was *femminicidio* (femicide). This choice was undoubtedly influenced by the urgent and increasingly shared need to finally consider the issue of gender-based violence as a pervasive structural problem at the socio-cultural level.

However, the question arises as to how one can effectively promote and raise awareness regarding the informed utilisation of said word and, consequently, the understanding of the gravity and systemic dimension of femicide, if the media, which should be the first to engage in this, pursue an unhealthy narrative with the sole aim of catching views to increase interaction and, therefore, revenue.

The present research aims to delineate, utilising purpose-built *NooJ* syntactic grammars, the dominant narrative patterns and strategies in the Italian press based on a dataset comprising 120 news articles from 2023 categorised under the tag *femminicidio*. All texts were published by three relevant newspapers, namely *Il Corriere della Sera*, *La Repubblica* and *Il Sole 24 Ore*, which were chosen as exemplars for the statistical survey due to their widespread national circulation, rendering them representative of the media landscape.

On the basis of the results, the goal of the proposed study is to highlight the points where a radical change in narrative methods would be needed in order to make the world of information and communication as transparent, objective and respectful as possible. Considering the urgency of the associated societal issues and the constantly changing methods and platforms of communication, such a research holds the potential to offer a concrete and solution-oriented outlook through an enhanced understanding of the commonly adopted narratives throughout the Italian newspapers.

Keywords

Femicide, gender-based violence, newspapers, syntactic grammars, analysis

References

1. Busso, L., Combei, C. R., & Tordini, O. (2020). The mediatization of femicide: a corpus-based study on the representation of gendered violence in Italian media. Analisi Linguistica e Letteraria, 28(3), 29-48.
2. Treccani (2023). "Femminicidio è la parola dell'anno 2023". In Treccani Magazine, 28th of December 2023. https://www.treccani.it/magazine/atlante/societa/femminicidio-e-la-parol a-dell-anno-2023.html (last accessed 12.04.2024)
3. Silberztein, M. (2003–). NooJ Manual. Available at: http://www.noojassociation.org
4. Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach. Wiley-ISTE, London.

Victimization and otherness: Unearthing colonial and post-colonial tropes in the EU migration policy arena using NooJ

Lorenzo Maggio Laquidara
Duke Kunshan University [1]
X23 Science in Society [2]
Kunshan, Suzhou, China [1]
Treviglio, Italy [2]
lorenzo.laquidara@dukekunshan.edu.cn
lorenzo.maggio@x-23.org

Davide Carminati
X23 Science in Society
Treviglio, Italy
davide.carminati@x-23.org

Abstract

Immigration management represents a central topic in policy agendas across modern-day E.U. countries. Official political rhetoric relies on a number of figures in portraying low-income immigrants as a policy issue. While anti-immigration voices construe foreign-born citizen inflow as an economic, societal, and cultural threat, echoing narratives of religious, racial, and ethnic competition. In contrast, political discourses advocating for migrant assistance and integration tend to espouse a victimizing narrative. Both rhetoric forms have been widely regarded as biased, and ultimately unfit to capture the complexity of this fragmented interregional socio-economic phenomenon. Migration scholarship has traced these biased narratives back to the departure and destination countries' colonial past and post-colonial present. Because European views on immigration reflect deep-rooted historical reciprocities, they transpire from political discourses as seeming common sense, making them often hard to spot and demand accountability for. This paper intends to reveal those narratives that resonate with Eurocentric colonial and post-colonial rhetoric in modern European political literature. NooJ-mediated discourse analysis will create token frequency tables that can be mapped onto the narrative cues common to biased migration policy talk. Relevant tokens will be compiled into conceptual clusters, which will hopefully represent the key logical strongholds of political discourse. The analysis's clustering is rooted in sociological concepts relevant to colonial and migration discourses. Achilles Mbembe's blackness ideas are investigated in the study as a key conceptual framework to investigate the interactions between im/migrant flows and the broad European society. Otherness-centered discourses will be contrasted with constructive considerations on social navigation, acculturation, and cultural citizenship. To this extent, this study intends not only to combine sociological theory with grounded theory and thematic analysis techniques but to also expand the scope to which those theoretical dogmas extend. In fact, im/migration represents a new field of investigation and practice in sociological, anthropological, and ethnographic studies. In conclusion, the authors reflect on the tropes revealed by discourse analysis from social psychology and public policy perspectives. Colonial discourses are contrasted with narratives that resolve the underlying tensions existing across citizenship, racial, ethnic, and historical boundaries. Critical engagement will be followed by matching policy issues and relevant discourses with expectedly effective public policy devices. Therefore, this study attempts to move some first steps in the complex interdisciplinary field of E.U. migration policy, chiefly relying on NooJ-mediated corpus text analysis, and drawing from a diverse arsenal of analytical techniques alongside.

Keywords

Immigration, Public Policy Evaluation, Otherness, Corpus Analysis

References

1. Mbembe, A. (2017). *Critique of black reason*. Duke University Press.
2. Balibar, E. (2013). Masses, classes, ideas: Studies on politics and philosophy before and after Marx. Routledge.
3. Moore, J. W. (2016). The rise of cheap nature.
4. Olivieri, C., Laquidara, L. M., & Semlali, A. (2023, May). Semantic Analysis of Migrants' Self-entrepreneurship Ecosystem Narratives. In International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ (pp. 176-187). Cham: Springer Nature Switzerland.
5. Nikielska-Sekula, K., & Desille, A. (2021). Visual methodology in migration studies: New possibilities, theoretical implications, and ethical questions (p. 343). Springer Nature.

SOCIAL SCIENCES APPLICATIONS

Unveiling Migrant Entrepreneurial Narratives: A Discourse Analysis Approach for Inclusive Support

Cecilia Olivieri

X23 Science in Society

Treviglio, Italy
cecilia.olivieri@x-23.org

Lorenzo Maggio Laquidara

Duke Kunshan University [1]
X23 Science in Society [2]

Kunshan, Suzhou, China [1]
Treviglio, Italy [2]
lorenzo.laquidara@dukekunshan.edu.cn
lorenzo.maggio@x-23.org

Agathe Semlali

X23 Science in Society

Treviglio, Italy
agathe.semlali@x-23.org

Davide Carminati

X23 Science in Society

Treviglio, Italy
davide.carminati@x-23.org

Abstract

Migrant entrepreneurs in Europe constitute a multifaceted group, represented by very diverse and complex identities, articulated in multiple generations, places of origin, demographic characteristics, and ambitions. The exploration of this target group is a crucial endeavour, given the unique challenges and opportunities it faces in exploring unfamiliar economic landscapes. In spite of its sociological, political, and civic relevance, approximative narratives on migrant entrepreneurship are widespread among institutional actors. Support organisations in entrepreneurship training often portray migrants as a vulnerable social group and bearers of basic needs only [5]. Other than running the risk of fostering the sense of alterity between locals and migrants, normally these narratives homologate the perception of migrants in society and disable a diversified support system. This is the reason why self-perception directly collected from the target group represents a valuable way to overcome top-down narratives and navigate migrants own perspectives around their socio-economic context [1-6], needs and dreams - all valuable elements that drive understanding about the real pathway that they intend to pursue and the actual support they need to achieve it.

In the present study we intend to implement the discourse analysis model elaborated by Olivieri et al., [3] based on co-occurrence analysis through NooJ [4] semantic grammars for the analysis of migrant entrepreneurs interviews. Such a model aims to go beyond statistical metrics and delve into the semantic spheres evoked by migrant entrepreneurs: by identifying the core themes that hold significance for the entrepreneurs and their interconnection, the research strives to reveal the underlying beliefs and inclinations of the speakers as they narrate their entrepreneurial journeys. By uncovering these connections, the present work aims to provide a deep and representative understanding of the speakers' needs, skills and aspirations, offering insights that can inform effective support strategies during the incubation process. The ultimate goal is to encourage more inclusive, empowering, and personal attuned support systems that enable migrant entrepreneurs to succeed in their entrepreneurial pathway.

Semi-structured interviews are the main instrument to explore the narratives of 10 selected migrants. Coherently with the identified model [3] the study intends to process these transcribed natural-language inputs using NooJ, relying on co-occurrence analysis of semantic and syntactic [2] grammars - namely concepts. This is to reveal which semantic fields are proposed by the speaker and how they interact between each other. The interaction between concepts is investigated by comparing the normal distribution of the different semantic and syntactic grammars along each text. By studying the co-occurrences between concepts and their distribution in the discourse, the most relevant themes for the speaker are identified (on a qualitative and non-statistical basis); potential relations between concepts are highlighted and a need analysis for a better and more tailored entrepreneurial support is explored.

Keywords

Migrant Self-Entrepreneurship, Formal Semantic Analysis, NooJ Grammars, Tailor Made Support Models, Incubation Programs

References

1. Buhr, F. (2021). Migrants Mental Maps: Unpacking Inhabitants Practical Knowledges in Lisbon. In: Nikielska-Sekula, K., Desille, A. (eds) Visual Methodology in Migration Studies, 51–67. Springer, Cham.
2. Monteleone, M. (2019). NooJ Grammars and Ethical Algorithms. Università degli Studi di Salerno, Fisciano.
3. Olivieri, C., Sheng, J., Laquidara, L. M., & Semlali, A. (2023). Semantic Analysis of Migrants' Self-Entrepreneurship Ecosystem Narratives. Proceedings of 17th International Conference, NooJ 2023, Zadar, Croatia, May 31- June 2, 2023. Springer International Publishing.
4. Silberztein, M., Monti, J., Monteleone, M., eds. (2014). Formalizing Natural Languages with NooJ. Cambridge Scholars Publishing, Newcastle upon Tyne.
5. Van Rijs S. B.; Scalera V.G., Lupoi O. (2021) RESEARCH REPORT. INCUBATION SERVICES FOR MIGRANT ENTREPRENEURSHIP MIG.EN.CUBE: Fostering MIGrant ENtrepreneurship inCUBation in Europe, University of Amsterdam (the Netherlands)
6. Zanfrini, L. (2015). The Diversity Value: How to Reinvent the European Approach to Immigration. McGraw-Hill Education, Maidenhead.

Beyond Sustainable Development, Unraveling Regenerative Tourism through NooJ

Lorenzo Maggio Laquidara

Duke Kunshan University [1]
X23 Science in Society [2]

Kunshan, Suzhou, China
Treviglio, Italy
lorenzo.laquidara@dukekunshan.edu.cn
lorenzo.maggio@x-23.org

Agathe Semlali

X23 Science in Society

Treviglio, Italy
agathe.semlali@x-23.org

Abstract

The tourism industry is changing as fast as society is. The industry is exposed to a plethora of challenges, subject to changes in tourist preferences, regulations and policies, climate change, and political and economic context. Tourism is a vector of socio-economic prosperity in a globalised world [2]. However, its impact is not without leaving a trace [5]. Despite significant economic benefits, tourism is aggressive toward the environment, local population, and culture. Local ecosystems are bearing the costs of over-tourism practices (e.g., Venice, Italy). Since the Agenda 2030 of the United Nations, the sector has experienced a rising trend of sustainable tourism practices and policies [3]. Indeed, whether from governmental, intergovernmental, or local initiatives, tourism stakeholders are trying to ensure that touristic practices comply with ethical and ecological standards (e.g., green tourism, slow-tourism, agrotourism, rural tourism, solidarity tourism, and more). However, while sustainable tourism aims to halt or minimise harm to ecosystems and socio-economic realities, new trends such as the regenerative tourism approach aim to repair, enhance, and heal the places we visit.

This research extracts and analyses texts from different tourism-related initiatives from both public and private entities; on one side, self-defined sustainable tourism initiatives and practices; on the other, self-proclaimed regenerative tourism initiatives and practices. In this approach, we include organisations working toward restorative and reparative tourism. This paper considers purposively regeneration, reparation, and restoration as synonyms - although the concept of regenerating goes a step further [1].

The data is constituted of extracts of advocacy and promotional text from a diverse range of organisations. They all work to promote sustainable or regenerative tourism practices or advocate for a mindset change in the industry and among tourists. The research focuses on the lexicon used in both approaches. Therefore, we used NooJ, a powerful Natural Language Processing tool [4]. The texts are processed through NooJ to grasp key concepts and recurring lemmas. To strengthen the approach, grammars are built for each paradigm. These grammars are reinforced throughout the process to create a precise recognition tool for future tourism texts.

The overall objective of this research is to understand and contrast the linguistic differences between the two approaches, sustainable and regenerative tourism, to then grasp their ideologic convergences or divergences. Therefore, the findings also enable the conceptualisation of the new paradigm of regenerative tourism to meet the challenges of contemporary tourism.

Keywords

Sustainable Tourism, Regenerative Tourism, NooJ, Discourse Analysis

References

1. Centre for the Promotion of Imports from developing countries (CBI). (2022, May 2). Regenerative tourism: Moving beyond sustainable and responsible tourism. CBI - Minister of Foreign Affairs The Netherlands. https://www.cbi.eu/market-information/tourism/regenerative-tourism
2. Franklin, A., and Crang, M. (2001). The trouble with tourism and travel theory? Tourist Studies, 1(1), 5-22. https://doi.org/10.1177/146879760100100101
3. Khizar, H. M. U., Younas, A., Kumar, S. , Akbar, A. and Poulova, P. (2023). The progression of sustainable development goals in tourism: A systematic literature review of past achievements and future promises. Journal of Innovation & Knowledge, 8(4).
4. Monteleone, M. (2019) NooJ Grammars and Ethical Algorithms. Università degli Studi di Salerno, Fisciano.
5. World Tourism Organization. (2023). Achieving the Sustainable Development Goals through Tourism ? Toolkit of Indicators for Projects (TIPs).

Syntactic Analysis of Media Reporting on Feminicides and Gender-Based Violence in Modern Italy

Giorgia Bozzetto

Università degli Studi di Salerno
Salerno, Italy
g.bozzetto@studenti.unisa.it

Nicola Cannoniero

Università degli Studi di Salerno
Salerno, Italy
n.cannoniero@studenti.unisa.it

Caterina Lausi

Università degli Studi di Salerno
Salerno, Italy
c.lausi@studenti.unisa.it

Stefano Madonna

Università degli Studi di Salerno
Salerno, Italy
s.madonna@studenti.unisa.it

Camilla Nastro

Università degli Studi di Salerno
Salerno, Italy
c.nastro7@studenti.unisa.it

Anna Maria Petruzzi

Università degli Studi di Salerno
Salerno, Italy
a.petruzzi1@studenti.unisa.it.

Abstract

Origin of the modern usage of the term "femicide", describing a killing motivated by misogyny, is generally attributed to Diana Russell and Jill Radford in their 1992 book on the subject [1]. The public at large arguably mostly familiarized with the term within the 2010-20 decade. In Italy the conversation around misogynistic gender-based violence received mainstream attention in that same period, with recent news and media coverage generating a new peak for the term “femminicidio” in 2023 Google Search stats [2].

Recent data tells of 120 femicides in 2023 Italy alone, with romantic/marital partners recognized as aggressors in more than half of the cases [3]. Using the NLP analysis tool NooJ, our research tries to capture the state of the media-led conversation around femicides, through the examination of reporting surrounding seven high-profile criminal cases that occurred from 2005 to 2023. Established writing on the subject of media narratives concerning gender violence such as the "Venice manifesto" [4] was consulted to inform the research on a corpus of more than 120 online articles. Six NooJ grammars were created to analyze frequencies and generate statistics around potentially misleading and misogynistic tropes occurring in the corpus, such as pseudo-apologetic attitudes displayed towards the aggressors, inattentive infantilization of the victims, emphasis placed on women’s role as mothers, wives and daughters, the recurrence of secondary victimization. We also considered the way modern tech and social media might influence both feminist advocacy campaigns and dynamics surrounding gender-based violence. The reconstruction of chronological events emerging from the corpus also highlighted a pattern of legal complaints and warning signs ultimately often ignored by authorities or society at large [5].

The results paint a picture where the discussion around misogyny as a potential motivator of homicidal violence has become mainstream despite the relative perseverance of distortions that might contribute to secondary victim blaming.

Keywords

Femicide, Gender-Based Violence, Syntactic Analysis, Italian Media Analysis

References

1. Radford & Russell, eds. (1992). Femicide: The Politics of Woman Killing, Open University Press.
2. Google Trends, "Femminicidio" (2004). Current day graph. https://trends.google.it/trends/explore?date=all&geo=IT&q=femminicidio&hl=it
3. Ministero dell'Interno, Dipartimento della Pubblica Sicurezza (2024). "Analisi criminologica della violenza di genere", https://www.interno.gov.it/sites/default/files/2024-02/elaborato_analisi_viol_gen.pdf
4. Sindacato Giornalisti Veneto. (2017). "Manifesto delle giornaliste e dei giornalisti per il rispetto e la parità di genere nell'informazione", https://www.sindacatogiornalistiveneto.it/wpcontent/uploads/2020/12/MANIFESTO-DI-VENEZIA.pdf
5. Commissione parlamentare di inchiesta sul femminicidio, nonché su ogni forma di violenza di genere. (2021). https://www.senato.it/service/PDF/PDFServer/DF/366054.pdf

Linguistic Analysis on Project Evaluation through NooJ: the EU WeMed_NaTour Pilot Trip Assessment

Agathe Semlali

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

Jiwon Shin

X23 Science in Society
Treviglio, Italy
jiwon.shin@x-23.org

Marika Mazzi Boém

X23 Industry
Treviglio, Italy
marika.mazzi.boem@x-23.org

Abstract

The EU WeMED_NaTOUR project, funded by the European Union, aims to develop transnational eco-smart tourism packages targeting coastal and natural marine destinations in the West Mediterranean. The goal is to enhance the attractiveness of less-known destinations, with a special focus on younger generations. The project wants to enhance stakeholders' capacity building by integrating sustainable tourism concepts into school trip curricula and developing comprehensive guides for sustainable practices. A pilot trip is an integral component of the project, involving practitioners in the tourism industry, beneficiaries from schools, and local SMEs.

This research investigates the effectiveness of the EU WeMed NaTour project by evaluating how well stakeholders implemented its goals. It also aims to measure the impact of the eco-educational trips on both teachers and students.

During the project, questionnaires are designed to collect meaningful insights from the three different targeted respondents: stakeholders, teachers, and students [2]. The collected data from the answers enables the evaluation of the project implementation to further fine-tune the product features based on the beneficiaries' desiderata. NooJ, as a robust Natural Language Processing software, allows to perform several NLP routines to intersect quantitative and qualitative analytical methodologies: word occurrence, text parsing, and semantic-syntactic analyses. Such routines enable one to perform a Sentiment Analysis characterised by a deeper comprehension of the different audiences needs and opinions [1].

The first step of our analysis will consist in cleaning our data, which will be subsequently processed through NooJ [3]. During this step, NLP procedures will be performed, such as word frequency and co-occurrences analysis. As known, co-occurrence analysis can highlight relationships between concepts, while semantic-syntactic analysis by means of grammars can unveil sentiments, attitudes and opinions. During a second step, results will be classified on the basis of their categories that is, respondents' groups and language used. Being questionnaires with short and moderate-length responses, texts are expected to require a low level of disambiguation.
Aside from its scientific purpose, the research will help produce an "Analytical matrix report" of optimized parameters for the project managers, while consolidating "Good practices" factors.

The outputs will allow for a thorough comprehension of the viewpoints held by all the stakeholders involved. The outcome of this study will be used to develop future eco-tourism strategies incorporating sustainability and the blue economy.

Keywords

Ecologic tourism, Reparative tourism, Project Evaluation, Questionnaire, NooJ

References

1. Bendinelli, M. (2022). Who is to Blame for What? An Insight Within the French Yellow Vests' Movement Through Dole's Books of Grievances. Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities: Revised Selected Papers from the NooJ 15th International Conference, NooJ 2021 (Besançon, France). Edited by Magali Bigey, Annabel Richeton, Max Silberztein, and Izabella Thomas. Springer Nature, Switzerland.
2. Bihu, R. (2021). Questionnaire Survey Methodology in Educational and Social Science Studies. International Journal of Quantitative and Qualitative Research Methods Vol.9, No.3, pp.40-60.
3. Silberztein, M. (2016) Formalizing Natural Languages: The NooJ Approach. Wiley-ISTE, London.
4. Zhu, J.J., Wang, H., Ye, C., and Lang, Q. (2015). Project evaluation method using non-formatted text information based on multi-granular linguistic labels, Information Fusion, 24, 93-107, https://doi.org/10.1016/j.inffus.2014.09.006.
5. * The European-funded project EU-WEMED-NATOUR website: https://www.euwemed-natour.eu/

INFO AT:

https://nooj2024.x-23.org







Printed at **Eliotecnica s.n.c.,** Bergamo - Treviglio, Italy on May 25, 2024.

Notes:

Notes: