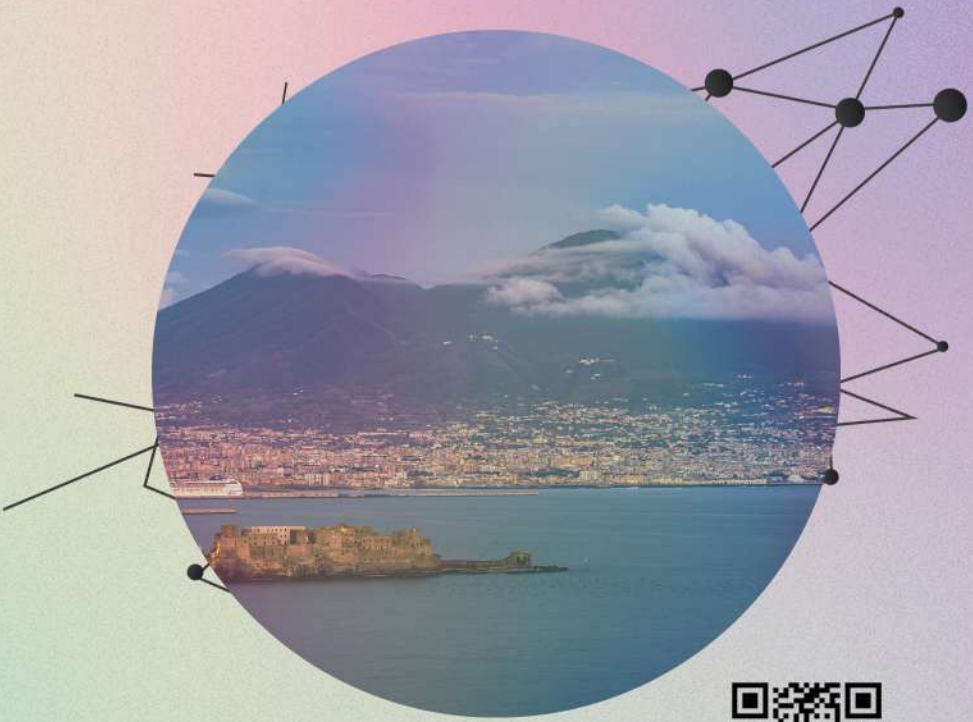
THE 20th NOOJ INTERNATIONAL CONFERENCE 2026

24-26 JUNE 2026

Palazzo Giusso
Largo San Giovanni Maggiore, 30
Aula Matteo Ripa



UNIVERSITÀ DI NAPOLI L'ORIENTALE

THE 20TH NOOJ INTERNATIONAL CONFERENCE 2026

Book of Abstracts

UNIVERSITÀ DI NAPOLI
L'ORIENTALE

24-26 June 2026
Naples - Italy

Table of Contents

Programme

Palazzo Giusso, Largo San Giovanni Maggiore, Aula Matteo Ripa

24 June 2026

9:00 - 9:40	**Registration**	
9:40 - 10:00	**Conference Opening**	
	Oral Session I - Natural Language Processing	
10:00 - 10:20	Bona-Pellissier Morgane	Modelling a Fusional Language: From NooJ Grammar to Python Parsing, the Case of Latin Conjugation
10:20-10:40	Meroni Fabio and Duran Maximiliano	Linguistically Grounded NLP for Low-Resource Languages: Parallel Experiences from Quechua and Mwotlap
10:40-11:00	**Coffee Break**	
	Oral Session II - Natural Language Processing	
11:00-11:20	Favilla Anna, Riunno Maria Victoria, Borrelli Giulia and Makarov Oleksander	From natural language use to automatic disambiguation of the italian connector “che”
11:20-11:40	Terrone Francesco, Rodrigo Andrea Fernanda and Enriquez Javier	NooJ as a Symbolic Interface: Recursive Grammars and Formal Linguistic Representation in Digital Humanities
11:40-12:00	Abbes Amira, Fehri héla, Zayani Corinne Amel, Ghorbel Leila and Champagnat Ronan	Process Mining Meets Linguistics: Discovering and Translating Learning Paths on Moodle with NooJ
12:00 - 12:20	Maisto Alessandro and Martorelli Giandomenico	Extracting metaphorical senses of verbs with selectional restriction with NooJ
12:20-12:40	Prihantoro Prihantoro	Natural Language Query Interface
12:40-13:00	Bigey Magali	From Media Panic to Moral Panic: The Gradation of Media Vocabulary

		Surrounding Dark Romance Analyzed with NooJ
13:00 - 14:20	**Lunch break**	
14:20-15:20	Tutorial	
15:20-16:00	**Coffee break**	
	Oral Session III - Natural Language Processing	
16:00-16:20	Suprunchuk Mikita, Hetsevich Yuras and Varanovich Valery	Modelling of the predicate structure of a sentence for Belarusian and English using NooJ
16:20- 16:40	Rodrigo Andrea Fernanda	A Corrector Prototype for Spanish (Corrector LLCIA_UNR)
16:40-17:00	Bešlić Katja	The Ball is Round: Idioms in Sports Journalism - SuperSport Croatian Football Cup
17:00-17:20	Duran Maximiliano	Automatic retrieval of Quechua adjectival forms
17:20 - 19:00	**Welcome cocktail**	

25 June 2026

9:00 - 9:40	**Registration**	
	Oral Session IV - Linguistic Resources	
9:40 - 10:00	Viktorija Borko	Warriors and Princesses: Gender Patterns of Metaphors in Croatian Sports Journalism
10:00 - 10:20	Silvio Calderaro and Johanna Monti	Cantada a bolu, Computational Analysis of Sardinian Extemporaneous Poetry
10:20-11:20	KEYNOTE: **Rachele Sprugnoli:** *Linking Linguistic Resources for Italian Dialects: Interoperability and Applications*	
11:20-12:00	**Coffee break**	
	Oral Session V - Linguistic Resources	
12:00 - 12:20	Kocijan Kristina and Farkaš Daša	Noun phrase complexity in a corpus of children's writing
12:20-12:40	Ritamaria Bucciarelli	Topological Braiding of Linguistic Structures: Anyons, Recursion and Symbolic Robustness in NooJ
12:40-13:00	Mijic Linda	Teaching Latin Morphology through Resource Building in NooJ
13:00-15:00	**Lunch break**	
	Poster Session	
15:00-16:30	Kristina Kocijan and Šojat Krešimir	Improving Croatian NLP Infrastructure: An Expanded and Structurally Annotated Verb Dictionary for NooJ
15:00-16:30	Faride Nikpour	Adapting the Arabic NooJ Module for Persian: A Pilot Study
15:00-16:30	Giulia Speranza	Naming the Hero: A Computational Study of Referential Variation of Achilles and Odysseus in Italian Homeric Translation
15:00-16:30	Maria Pia di Buono and Mariapia Battipaglia	Say It Again: A Semantic Rule-Based Taxonomy of Medical Term Simplification in Italian Patient Leaflets
17:00-19:00	**City Tour**	
19:30	**Social Dinner** *Antonio&Antonio,* Via Partenope, 26, 80121 Napoli NA	

26 June 2026

9:30 - 10:00	**Registration**	
	Oral Session VI - Linguistic Resources	
10:00 - 10:20	Max Silberztein	Grammatical Words and Stop Words
10:20-10:40	Rabiei Marzieh	Linguistic and Morphological Analysis of Persian Nouns: Utilizing the NooJ Platform
10:40 - 11:00	Kaliska Agnieszka	A Computational Approach to Optimize Verbal Inflectional Resources with NooJ
11:00-11:40	**Coffee break**	
11:40-12:00	Nougaret Lara	From coded lexicon to NooJ dictionary: property-based annotation of ecological lexical innovation
12:00 - 12:20	Rabiet Victor and Petković Divna	Interactive Web Interface for Learning Ukrainian with NooJ
12:20 - 14:20	**Lunch break**	
	Oral Session VII - Digital Humanities	
14:20-14:40	Joubert Maïa	The Discourse of Dubbing: Balancing Art and Technique in Professional Identity – A NooJ-Aided Analysis
14:40-15:00	Santonicola Alessia, Santarsiere Rosa and Capitelli Francesco Ugo	The Representation of the Black Lives Matter Movement in US Media: A Comparative Lexical Analysis
15:00-15:30	**Coffee break**	
15:30-16:00	**General NooJ Assembly**	
16:00-16:20	**Closing Remarks**	

Organizing Committee & Conference Chairs

Max Silberztein, Université de Franche-Comté, France
Johanna Monti, Università di Napoli L'Orientale, Italy
Maria Pia di Buono, Università di Napoli L'Orientale, Italy

Scientific Committee

Marco Angster, University of Zadar, Croatia
Anabela Barreiro, INESC-ID, Portugal
Anita Bartulović, University of Zadar, Croatia
Magali Bigey, Université de Franche-Comté, France
Xavier Blanco, Autonomous University of Barcelona, Spain
Christian Boitet, Université Grenoble Alpes, Grenoble, France
Maria Pia di Buono, Università di Napoli L'Orientale, Italy
Outi Duvallon, L'Institut national des langues et civilisations orientales (INALCO), France
Héla Fehri, University of Sfax, Tunisia
Zoe Gavriilidou, Democritus University of Thrace, Greece
Yuras Hetsevich, National Academy of Sciences, Belarus
Agata Jackievicz, Université Paul Valéry, France
Agnieszka Kaliska, Poznan University, Poland
Kristina Kocijan, University of Zagreb, Croatia
Walter Koza, National University of General Sarmiento, Argentina
Svetlana Krylosova L'Institut national des langues et civilisations orientales (INALCO), France
Mathieu Lafourcade, Université de Montpellier, France
Laetitia Leonarduzzi, Université d'Aix-Marseille, France
Samir Mbarki, IbnTofail University, Morocco
Johanna Monti, Università di Napoli L'Orientale, Italy
Kamal Naït-Zerrad, L'Institut national des langues et civilisations orientales (INALCO), France
Thierry Poibeau, Laboratoire Lattice, CNRS, France
Andrea Rodrigo, University of Rosario, Argentina
Olena Saint-Joanis, L'Institut national des langues et civilisations orientales (INALCO), France
Max Silberztein,Université de Franche-Comté, France
Marko Tadić, University of Zagreb, Croatia
François Trouilleux, Université Clermont Auvergne, France

Keynote Speaker

Rachele Sprugnoli, Università Cattolica del Sacro Cuore, Italy

Local Support Team

Mariapia Battipaglia, Università di Pisa, Università di Napoli L'Orientale, Italy
Silvio Calderaro, Università di Pisa, Università di Napoli L'Orientale, Italy
Petra Giommarelli, Università di Pisa, Università di Napoli L'Orientale, Italy

Message from the Organising Committee

This volume contains the abstracts of the contributions to the *20th NooJ International Conference (NooJ 2026)*, held at the University of Naples "L'Orientale", Italy, from 24 to 26 June 2026. The conference marks the twentieth edition of the annual international meeting dedicated to NooJ, a linguistic development environment and corpus processing platform that enables the formalization of linguistic phenomena at orthographic, lexical, morphological, syntactic, and semantic levels, and supports the development of a wide range of Natural Language Processing (NLP) applications.

Over the past two decades, the NooJ conference series has established itself as a major forum for researchers, developers, educators, and practitioners working at the intersection of Linguistics, Computational Linguistics, Digital Humanities, and Language Technologies. Beyond presenting advances in the NooJ platform itself, the conference has fostered an international community dedicated to the development of linguistic resources and the exploration of language-based computational approaches across a wide variety of languages and research contexts.

The 2026 edition continues this tradition by bringing together scholars from different countries and disciplines to exchange ideas, methodologies, and experiences related to the formal description and computational processing of natural languages. The conference provides a venue for discussing both theoretical and applied research, with particular attention to lexical resources, computational morphology, syntactic and semantic analysis, and linguistically motivated NLP applications.

The contributions collected in this volume reflect the diversity of current research conducted with and around NooJ. They present new linguistic resources, innovative methodologies for corpus analysis, advances in computational modeling, and applications addressing challenges in language processing, digital humanities, language teaching, information extraction, machine translation, and semantic annotation. Together, these studies demonstrate the continuing relevance of linguistic formalisms and rule-based approaches within the rapidly evolving landscape of language technologies and artificial intelligence.

In addition to the presentation of research results, NooJ 2026 also aims to strengthen collaboration among members of the international NooJ

community and to support the training of new generations of researchers. Following a long-standing tradition of the conference series, the programme includes opportunities for discussion, methodological exchange, and advanced training in the automatic analysis and generation of texts.

We would like to express our sincere gratitude to all authors for submitting their work, to the members of the Scientific Committee for their careful evaluation of the contributions, and to the Department of Literary, Linguistic and Comparative Studies of the University of Naples “L’Orientale” for supporting the conference. Our thanks also go to the wider NooJ community, whose commitment and collaborative spirit have contributed to the success and longevity of this international conference series.

We hope that readers will find these abstracts both informative and inspiring, and that the contributions presented here will stimulate further research and collaboration in Linguistics, Computational Linguistics, Digital Humanities, and language technology development.

Naples, June 2026

Max Silberztein, Johanna Monti, and Maria Pia di Buono

Invited Speaker

Rachele Sprugnoli
Università Cattolica del Sacro Cuore

Rachele Sprugnoli is a computational linguist and digital humanist. She obtained her bachelor's and master's degrees in Humanities Computing at the University of Pisa and her PhD in Information Technology at the University of Trento. She worked at Fondazione Bruno Kessler in Trento and at the University of Parma. Currently, she is consultant for NLP projects at the Università Cattolica del Sacro Cuore in Milan and adjunct professor at the University of Torino. Her research focuses on the application of computational approaches to the study of historical and literary texts in Italian, Latin and English.

Linking Linguistic Resources for Italian Dialects: Interoperability and Applications

This talk explores how linguistic resources for Italian dialects can be modeled and interlinked as Linguistic Linked Open Data (LLOD) to enhance interoperability, reuse, and integration. Dialectal data, often fragmented across lexicons, corpora, and grammars, can be transformed into structured, ontology-driven knowledge graphs using shared standards. By linking these resources to existing infrastructures, this approach enables cross-resource navigation and comparative analysis. Looking ahead, the talk reflects on how such interoperable knowledge may shape future AI applications.

Modelling a Fusional Language: From NooJ Grammar to Python Parsing, the Case of Latin Conjugation

Morgane Bona-Pellissier
Université Paris Nanterre, 92000 Nanterre, France

Latin, as a prototypical fusional language characterized by high cumulation [1] and syncretism [2], presents distinct challenges for computational morphological modelling. This paper discusses the hurdles encountered when modelling Latin conjugation with NooJ [3] and proposes a hybrid workflow combining a specific NooJ encoding strategy with a custom Python postprocessing library, pynooj [4].
The primary challenge lies in the "feature conflict" arising from the recursive reuse of morphemes, which renders standard compositional "Item and Arrangement" approach [5, p. 212] inadequate. For example, identical morphemes often carry contradictory tags in different contexts, such as the pluperfect indicative (*amaueram*), formed using the marker "-era-", which is formally identical to the imperfect form of esse (*eram*) [6, p. 590]. To resolve such conflicts, we adopt a theoretical framework based on Aronoff's concept of "Priscian" formations [7], implementing a systematic "double encoding" strategy. Ambiguous morphemes twice: once as lexical entries, and again as "whitewashed" morphological markers. Consequently, our model aligns with the "Word and Paradigm" framework [8], prioritizing the integrity of the inflectional cell over linear concatenation.
Simultaneously, we address local irregularities through deliberate "overgeneration", generating both the correct irregular form and the theoretical regular "false positive" to maintain graph simplicity. To operationalize this strategy, *pynooj* converts NooJ's dictionary output into structured JSON objects containing rich linguistic metadata. This Python layer serves as a dual-validation quality control instrument: it ensures high recall by verifying paradigm completeness across every cell, and maximizes precision by filtering out the over-generated false positives. Finally, the workflow integrates an AI component to automatically map each inflected form to its French translation, effectively bridging the gap between morphological generation and semantic interpretation.

Keywords: Latin Morphology, NooJ, Python, Fusional Languages, Priscian Formations, Word and Paradigm

References

[1] P. H. Matthews. Inflectional Morphology: A Theoretical Study Based on Aspects of Latin Verb Conjugation. Cambridge Studies in Linguistics, 6. Cambridge University Press, 1972.
[2] J. Šnajder. Models for predicting the inflectional paradigm of Croatian words. Slovenščina, 2.0, 1(2):1–34, 2013.
[3] M. Silberztein. Formalizing Natural Languages: The NooJ Approach. ISTE-Wiley, UK, 2016.
[4] M. Bona-Pellissier. *pynooj*: Python library for parsing NooJ's dictionary files. Python Package Index. https://pypi.org/project/pynooj/. Accessed May 13, 2026.
[5] C. F. Hockett. Two Models of Grammatical Description. WORD, 10(2-3):210–234, 1954. https://doi.org/10.1080/00437956.1954.11659524.
[6] A. Sihler. New Comparative Grammar of Greek and Latin. Oxford University Press, UK.
[7] M. Aronoff. Morphology by Itself: Stems and Inflectional Classes. Linguistic Inquiry Monograph, 22. MIT Press, 1994.
[8] J. P. Blevins. Word and Paradigm Morphology. Oxford University Press, UK, 2016.

Linguistically Grounded NLP for Low-Resource Languages: Parallel Experiences from Quechua and Mwotlap

Maximiliano Duran[1,2] and **Fabio Meroni**[3]
[1] Université de Franche Comté (CRIT). Besançon, France
[2] Université de Grenoble (LIGA). Grenoble, France
[3] Universidad del País Vasco (UPV/EHU). Vitoria Gasteiz, Spain

More and more actively spoken Indigenous and minority languages are today well-documented, yet remain absent from the landscape of computational linguistics. This gap prevents researchers, communities, and educators from leveraging corpus-based approaches that have become standard for better-resourced languages. This joint contribution presents two independent but convergent initiatives, the development of NooJ modules for Quechua [1] and Mwotlap [2]. Although these languages belong to unrelated families and distinct geographical areas, they shared a similar profile before the advent of their respective NooJ linguistic resources: rich descriptive grammars and dictionaries together with corpora of a good amount of tokens, and at the same time an almost complete lack of tools for automatic analysis in the light of Natural Language Processing (NLP).

Working within the NooJ environment [3,4], each project adapts linguistic descriptions into formalised, interpretable, and scalable computational resources. For Quechua, this involved the construction of electronic dictionaries [5], modelling agglutinative and polysynthetic morphology (Duran, 2014), encoding productive derivation [1,6] and syntactic patterns [7] all based upon previous descriptive accounts such as the ones by Guardia Mayorga [8]. For Mwotlap, the module translates the extensive documentation by Alexandre François in terms of grammar [9] and lexicon [10] into electronic dictionaries, inflectional and derivational grammars, morphophonological rules, and syntactic parsers capable of handling phenomena such as serial verbs, complex predicates, reduplication, and the language's elaborate TAM system [2,11]. Both efforts combine well-established descriptive work with formal language processing, implementing formal grammars pertaining to various levels of Chomsky-Schützenberger's [12] hierarchy to generate linguistic resources that are transparent to linguists and functional for text and corpus exploration.

By placing these projects side by side, the presentation highlights a shared methodological framework grounded in the Boasian trilogy [13,14]. It also

reflects on the challenges and opportunities involved in building NLP tools for low-resource languages without relying on large datasets or black-box models based on statistical methods, Machine Learning (ML) or Artificial Intelligence (AI) [15,16]. The Quechua and Mwotlap modules demonstrate how linguistically informed approaches can help the analysis documented but digitally invisible languages, enabling richer linguistic research.

Keywords: NooJ, Quechua, Mwotlap, Low resource languages

References

[1] Duran M. 2017. Quechua module for NooJ multilingual linguistic resources for MT. In Automatic Processing of Natural-Language Electronic Texts with NooJ. NooJ 2016, pages 48–63. Springer.
[2] Meroni F. 2026. Introducing Mwotlap Language Resources for NooJ. In Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2025, pages 3–15. Springer.
[3] Silberztein M. 2004. NooJ: A cooperative, object-oriented architecture for NLP. In INTEX pour la linguistique et le traitement automatique des langues, pages 351–361. Presses Universitaires de Franche-Comté.
[4] Silberztein M. 2016. Formalizing natural languages: The NooJ approach. John Wiley & Sons.
[5] Duran M. 2017. Dictionnaire électronique français-quechua des verbes pour le TAL. Doctoral thesis, Université de Franche-Comté.
[6] Duran M. 2016. The annotation of compound suffixation structure of Quechua verbs. Automatic processing of natural-language electronic texts with NooJ. NooJ 2015, pages 29–40. Springer.
[7] Duran M. 2020. Transformation and paraphrases for Quechua sentiment predicates. In Formalizing natural languages: Applications to natural language processing and digital humanities. NooJ 2024, pages 55–65. Springer.
[8] Mayorga C.A.G. 1973. Gramática kechwa. Ediciones Los Andes.
[9] François A. 2001. Contraintes de structures et liberté dans l'organisation du discours: Une description du Mwotlap, langue océanienne du Vanuatu. Doctoral thesis, Université Paris-IV Sorbonne.
[10] François A. 2026. A Mwotlap–English–French cultural dictionary. http://tiny.cc/Mwotlap-dict. Accessed May 05, 2026.
[11] Meroni F and François A. 2025. Capturing constructional patterns in texts: A NooJ approach to the Mwotlap corpus. Presentation at the 13th Conference on Oceanic Linguistics (COOL). Canberra, Australian National University, June 24, 2025.

[12] Chomsky N. and Schützenberger M.P. 1963. The algebraic theory of context free languages. In Computer programming and formal systems, pages 118–161. North Holland.
[13] Evans N. and Dench A. 2006. Introduction: Catching language. In Catching language: The standing challenge of grammar writing, pages 1–39. De Gruyter-Mouton.
[14] Evans N. 2014. How to write a grammar of an undescribed language. https://llacan.cnrs.fr/fichiers/cours/Evans/grammar_undescribed_language.pdf. Accessed May 05, 2026.
[15] Silberztein M. 2024. Preface. In Linguistic resources for natural language processing, pages xviii–xxvi. Springer Nature.
[16] Meroni F. 2025. Bringing NL back with P: Defending linguistic methods in NLP for future AI applications. In Proceedings of the 17th International Conference on Agents and Artificial Intelligence, volume III, pages 1062–1068. Scitepress.

From natural language use to automatic disambiguation of the italian connector “che”

Anna Favilla, Maria Victoria Riunno, Giulia Borrelli, Oleksandr Makarov
Università di Napoli “L’Orientale”, Naples, Italy

This study investigates the syntactic and semantic polyfunctionality of the Italian connector che, a highly frequent and multifunctional element that plays a central role in the organization of simple and complex sentences. While native speakers tend to process its different functions automatically and implicitly, its functional ambiguity represents a significant challenge for learners of Italian as a second language (L2), particularly in text comprehension and production across different registers.
The main goal is the systematic identification, formalization, and automatic disambiguation of selected grammatical functions of che through a rule-based computational approach.
The analysis is carried out using NooJ. A purpose-built corpus (CheAmb-ITA) of contemporary Italian texts (2018–2025) was compiled to ensure diafasic and diamesic variation, including informal written texts from social media, transcribed spoken language, and formal journalistic and academic sources, for a total of 26,871 tokens and 883 occurrences of che. All occurrences were manually annotated as a gold standard.
Local grammars were designed to automatically recognize functions such as relative pronoun (subject and object), interrogative and exclamative pronoun, and declarative and consecutive subordinating conjunction, then iteratively tested and refined to improve precision and reduce misclassification.
The results demonstrate satisfactory performance, particularly in the recognition of structurally stable functions. The study confirms the effectiveness and transparency of a rule-based approach for modeling grammatical ambiguity, highlighting its relevance for the development of computational linguistic resources and for potential applications in language teaching and advanced textual analysis.

Keywords: Rule-based disambiguation, Connector che, NooJ

References

[1] Berruto, G. Sociolinguistica dell’italiano contemporaneo. Carocci, Roma, 1998

[2] Corino, E. and Marello, C. Italiano di apprendenti. I corpora VALICO e VINCA. Guerra, Perugia, 2009
[3] Dardano, M. and Trifone, P. La lingua italiana. Zanichelli, Bologna, 1985
[4] Fiorentino, G. Che polivalente. In Enciclopedia dell'Italiano Treccani. Istituto della Enciclopedia Italiana, Roma, 2010
[5] Giacalone Ramat, A. (ed.) Verso l'italiano. Carocci, Roma, 2003
[6] Madsen, A., Reddy, S. and Chandar, S. Post-hoc interpretability for neural NLP: A survey. ACM Computing Surveys, 55(8), 1–42, 2022. https://doi.org/10.1145/3546577
[7] Renzi, L., Salvi, G. and Cardinaletti, A. (eds.) Grande grammatica italiana di consultazione. Il Mulino, Bologna, 2001
[8] Silberztein, M. NooJ Manual v7. Available for download at: https://nooj.univ-fcomte.fr
[9] Silberztein, M. NooJ: a Linguistic Annotation System for Corpus Processing. Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, Vancouver, 2005
[10] Zanda, F., Scerra, A., Santucci, V., Forti, L., Fioravanti, I., & Spina, S. Il Corpus Celi: Una Nuova Risorsa Per Studiare L'acquisizione Dell'italiano L2. *Italiano LinguaDue*, 2022
[11] Vietri, S. The Italian module for NooJ. In Proceedings of CLiC-it 2014, Pisa, 2014

NooJ as a Symbolic Interface: Recursive Grammars and Formal Linguistic Representation in Digital Humanities

Francesco Terrone [1], Ritamaria Bucciarelli [2], Andrea Fernanda Rodrigo [3] and Javier Julian Enriquez [4]
[1] Fondazione Francesco Terrone di Ripacandida e Ginestra, Italy
[2] Università Ca Foscari di Venezia, Italy
[3] Universidad Nacional de Rosario, Argentina
[4] Universitat Politècnica de València (UPV), Spain

This paper presents NooJ as a symbolic interface for formally representing of recursive linguistic structures in Digital Humanities and computational linguistics. The study argues that controlled grammar plays a central role in preserving interpretability, structural coherence, and semantic transparency. This stands in contrast to purely statistical or black-box approaches in Natural Language Processing.
Focusing on recursive linguistic mechanisms - such as repetition, anaphoric reference, and structural iteration - the paper illustrates how NooJ grammars enable explicit and computable representations of linguistic regularities that are particularly relevant for literary and poetic corpora. Through formal grammar and rule-based constraints, language is treated as a structured, symbolic system rather than opaque, probabilistic output.
The approach is grounded in a broader theoretical framework that considers recursion as a foundational principle shared by language, formal systems, and symbolic computation. Related applications to poetic language, including the formal analysis of Baudelaire's texts within Michel Planat's theoretical perspective, demonstrate the compatibility between symbolic grammar and advanced structural models.
In this context, NooJ operates as an operational environment that can embody abstract models through concrete grammatical implementations.
Presented by Francesco Terrone, this contribution provides a methodological counterpart to theoretical models of linguistic recursion, showing how NooJ grammars function as stable symbolic constraints for the formal analysis of complex linguistic structures in Digital Humanities.

Keywords: NooJ; symbolic grammar; recursion; formal linguistic representation; Digital Humanities; rule-based NLP; poetic corpora.

References

[1] M. Atiyah. The Geometry and Physics of Knots. Cambridge University Press, 1990.
[2] R. Bucciarelli. Topological Structures of Language: Rhetoric and Symbolic Computation. Archivio della Ricerca Ca' Foscari University of Venice (ARCA), 2025.
[3] R. Bucciarelli. Topology of the Word: Recursion and Emotional Computation. Ca' Foscari University of Venice, 2025.
[4] R. Bucciarelli. Topologies of the Mind. Language, Anyons, and Symbolic Consciousness. HAL Open Science, 2025.
[5] R. Bucciarelli. PROJECT: Language and Communication. QUIFL Project, Ca' Foscari University of Venice, 2025.
[6] N. Chomsky. Syntactic Structures. Mouton, 1957.
[7] M. Freedman, A. Kitaev, M. Larsen, and Z. Wang. Topological Quantum Computation. Bulletin of the American Mathematical Society, 40(1):31–38, 2003.
[8] D. R. Hofstadter. Gödel, Escher, Bach: An Eternal Golden Braid. Basic Books, 1999.
[9] D. Jurafsky and J. H. Martin. Speech and Language Processing. 3rd edn. Draft version, 2023.
[10] L. H. Kauffman. Knots and Physics. World Scientific, 2001.
[11] G. Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv preprint arXiv:2002.06177, 2020.
[12] J. Pearl. The Book of Why: The New Science of Cause and Effect. Basic Books, 2018.
[13] M. Planat and M. Amaral. What ChatGPT has to say about its topological structure: the anyon hypothesis. Machine Learning and Knowledge Extraction, 6(4):2876–2891, 2024.
[14] M. Planat, R. Aschheim, M. M. Amaral, F. Fang, and K. Irwin. Graph Coverings for Investigating Non-Local Structures in Proteins, Music and Poems. Sci, 3(4):39, 2021.
[15] M. Silberztein. Formalizing Natural Languages: The NooJ Approach. ISTE/Wiley, 2016.
[16] F. Terrone, R. Bucciarelli, A. F. Rodrigo, and J. Julian Enriquez. Gödel's Legacy: Formal Thinking, LLM and the Evolution of Artificial Intelligence. HAL Open Science, 2025.
[17] A. M. Turing. Computing machinery and intelligence (1950). Mind, 59(236):33–60, 1987.

Process Mining Meets Linguistics: Discovering and Translating Learning Paths on Moodle with NooJ

Amira Abbes[1,2], **Héla Fehri**[1], **Corinne Amel Zayani**[1], **Leila Ghorbel**[1], **Ronan Champagnat**[2] and **Jacques Morcos**[2]
[1] MIRACL LABORATORY, Pôle technologique de Sfax, France
[2] L3i Laboratory, La Rochelle University, Avenue Michel Crépeau, France

The use of Learning Management Systems (LMS) has significantly transformed higher education. These platforms record a wide range of student interactions and generate large amounts of data that can be analyzed to better understand learning behaviors and support personalized learning [1]. Several approaches exist for analyzing data from online learning systems. Among them, Process Mining has proven particularly effective in reconstructing, modeling, and analyzing real student learning paths from event logs [2]. Some studies propose recommending learning paths in the form of process models that students can follow, highlighting the potential of Process Mining to guide learners [3]. However, although informative, these models are often complex and difficult for students to understand. Improving the readability and interpretability of process models therefore remains an important challenge.
This paper presents a method that combines Process Mining [4] and the NooJ linguistic platform [5] to recommend comprehensible learning paths to students. This work extends our previous study [6], in which NooJ was used to classify student learning paths and provide recommendations through an interactive dashboard. In the proposed approach, we reuse the same dataset, including Moodle event logs and student grades. First, Process Mining techniques are applied to discover learning process models. Then, based on students' academic performance, the most successful learning paths are identified and recommended to other learners. To make the discovered process models understandable, NooJ transducers are used to translate these models into clear and readable textual descriptions.

This approach enhances the interpretability, transparency, and educational impact of learning analytics results, supporting a more student-centered learning environment.

Keywords: LMS, transducer, Process mining, learning analytics.

References

[1] Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. International Journal of STEM Education, 9(9).

[2] Cervantes, A. A., Mendoza, A., & Abedin, E. (2023). Uncovering Students' Learning Pathways: A Process Mining Perspective. In Proceedings of the 34th Australasian Association for Engineering Education (AAEE) Conference (pp. 746–754). Engineers Australia.

[3] Hachicha, W. (2024). Apport de la fouille de processus pour recommander des parcours d'apprentissage dans une plateforme pédagogique (Thèse de doctorat). Université de La Rochelle & Université de Sfax. https://theses.fr/s395378

[4] Van der Aalst, W. (2016). Process mining: Data science in action. Springer.

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

[6] Abbes, A., Fehri, H., Zayani, C. A., Ghorbel, L., & Champagnat, R. (2025). Personalizing Learning Paths: Analyzing Student Interactions on Moodle Using NooJ platform

Extracting metaphorical senses of verbs with selectional restriction with NooJ

Alessandro Maisto[1] and **Giandomenico Martorelli**
[1] University of Salerno, Fisciano, Italy

NOOJ [2] is a rule-based framework founded on finite state formalisms and explicit linguistic descriptions. In various contexts, it has been evaluated in terms of practical performance for natural language processing or to verify the industrial applicability of its capabilities. Nevertheless, NOOJ proved particularly suitable for the empirical comparison of lexicon-grammar (LG) descriptions involving large corpora.
In this work, we will adopt this perspective, using Nooj to test a theoretical hypothesis about the classes of verbs with selectional restrictions, as collected and described within the theoretical framework of the Lexicon-Grammar Theory (LG). In particular, we focused on classes 20R, which includes transitive verbs with restricted direct objects (e.g., *smoke, drink, wear, celebrate*), and 2B, of intransitive verbs with restricted subjects (e.g., *bark, growl, bloom*) [1].
We hypothesize that the more restricted their selection is, the more these verbs are characterized by a metaphorical use. Some examples of such uses concern expressions such as: *Fumare il cervello* (*to smoke the brain - mess with someone's head, blow someone's mind*), *Bere informazioni* (*to drink information - soak up information, absorb information*). Furthermore, the syntactic properties of the verbs change when they are used in figurative senses ([5,3]):

1. *Le galline covano le uova (The chickens are hatching eggs)*
2. *Le galline fanno una covata di uova (Chickens "make a hatching" of eggs)*
3. *Antonio cova* [*rancore, l'influenza, un virus*] (*Antonio is hatching* [*resentment, the flu, a virus*])
4. **Antonio fa una covata di* [*rancore, virus*] (**Antonio "makes a hatching" of* [*resentment, virus*])

The LG descriptions of those verbs reported in NOOJ through local lexicons and Finite State Grammars evidence how distributional violations openly modify operator-argument structures. The results can be compared with the concordances to observe the constructs that were validated and those that are marginal or not attested in actual usage. Once the

constructions with the selected verbs are collected, we exploit the dictionaries of the Italian Module [4] to test our hypothesis.
Our methodology is the following:

1. Local grammars are constructed to formalize the argument restrictions and syntactic alternations associated with each verb. The aim is to extract verb occurrences from the reference corpora in a verifiable, non-probabilistic way.
2. The extracted occurrences are analyzed to identify and compare the syntactic variations that distinguish concrete uses from figurative ones. The analysis focuses on qualitative aspects and quantitative/distributional aspects. The objective is to estimate the relative incidence of figurative versus concrete meanings for each verb and class.
3. The results are presented in terms of the distribution of verb occurrences across different semantic uses, distinguishing the number of concrete and figurative uses and generating a proportion. These proportions are correlated with a degree of syntactic-semantic restriction of the verb. This allowed us to compare different verb classes based on their concreteness and figurativeness and make a proportion between the two uses.

In conclusion, the analysis shows that NOOJ is a suitable tool for this type of lexicon-grammatical investigation. The use of dictionaries allows the construction of targeted lexical lists for non-probabilistic distributional analysis, while grammars allow a rapid and controlled data exploration, reducing the errors that tend to emerge in approaches based exclusively on probabilistic methods.
The paper highlights how NOOJ allows for a comprehensive analysis of syntactic hypotheses, rather than focusing on quantitative and performance metrics.
Usually, in this field, the absence of automatic analysis or probabilistic disambiguation is seen as a limitation. In this particular context, it can be a methodological advantage: certain linguistic assumptions must be made explicit so that their consequences can be observed and discussed directly.

Keywords: Verbs with Selectional Restrictions, Figurative Language, Lexicon Grammar.

References

[1] A. Elia, E. D'Agostino, M. Martinelli, et al. Lessico e strutture della sintassi. Introduzione alla sintassi del verbo italiano. 1981.

[2] M. Silberztein. Formalizing natural languages: The NooJ approach. John Wiley & Sons, 2016.
[3] S. Vietri. Idiomatic constructions in Italian. 2014.
[4] S. Vietri. The Italian module for nooj. In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa, pages 389–393. Pisa University Press, 2014.
[5] S. Vietri et al. Lessico-grammatica dell'italiano. UTET, 2004.

Connecting NooJ Grammars to CORTEX's Natural Language Query Interface

Prihantoro
Universitas Diponegoro, Semarang City, Indonesia

Formulating corpus queries often requires familiarity with formal query languages, posing a significant barrier for novice users [1, 2]. CORTEX, an advanced corpus analysis platform, supports a wide range of linguistic queries but, like many corpus tools, expects users to construct queries in a specific formal syntax (e.g. *ing _JJ _NN to retrieve words ending in *-ing* followed by an adjective and a noun). This paper presents an approach to bridge this gap by connecting NooJ grammars [3] to CORTEX's query engine in order to translate users' natural language queries into executable corpus queries. We design and implement a set of NooJ grammars that capture common linguistic query patterns expressed in natural language. Keywords and constructions such as *ending in*, *followed by*, *and*, as well as grammatical categories like *adjective* and *noun*, are modeled in NooJ to map user input to CORTEX's internal query format. In the proposed architecture, users submit natural language queries through the CORTEX interface, which are then sent to NooJ for grammatical analysis and translation before being executed by CORTEX. This rule-based approach builds on earlier work on grammar-driven query interpretation [4, 5]. To evaluate the approach, we compare NooJ-based translation with a lightweight large language model (Ollama tiny), reflecting recent trends in applying LLMs to natural language interfaces for data access [6]. The results show that NooJ outperforms the LLM on structurally complex queries, while both approaches perform similarly on simple keyword-based searches. In terms of processing speed, transparency, and determinism, the NooJ solution demonstrates clear advantages. We conclude that integrating NooJ grammars into corpus query systems is an effective and efficient strategy to lower the entry barrier for users and to support accurate natural language querying in corpus linguistics.

Keywords:
NooJ, CORTEX, Natural Language Query, Corpus Linguistics, Grammar-Based Parsing, Query Translation, Natural Language Interface, Corpus Query Language, Rule-Based Nlp, Large Language Models.

References

[1] J. Bernard and S. Evert. Quantitative corpus linguistics with R: A practical introduction. Springer, 2013.
[2] T. McEnery and A. Hardie. Corpus linguistics: Method, theory and practice. Cambridge University Press, 2012.
[3] M. Silberztein. Formalizing natural languages: The NooJ approach. ISTE / Wiley, 2016.
[4] M. Rospocher, L. Serafini, and M. Tesconi. From natural language questions to SPARQL queries: A survey of the state of the art. Semantic Web Journal, 7(1):1–20, 2016.
[5] B. Coppola, L. Laera, and M. Manna. Natural language interfaces to databases: A survey. Information Systems, 81:1–20, 2019.
[6] X. Li, Y. Zhang, J. Sun, and W. Zhou. Large language models for natural language to SQL: A survey. ACM Computing Surveys, 56(3):1–38, 2023.

From Media Panic to Moral Panic: The Gradation of Media Vocabulary Surrounding Dark Romance Analyzed with NooJ

Magali Bigey
Université Marie et Louis Pasteur, France
Affiliated Researcher CARISM Paris-Panthéon-Assas, France

For several years, Dark Romance has been the object of increased media coverage, linked to its strong visibility on social media platforms and to the expansion of its readership. This exposure is accompanied by a set of recurring media discourses, often alarmist, which tend to construct the genre as a problematic cultural object. The words mobilized to speak about it thus take part in a progressive discursive dynamic, in which media panic is transformed, step by step, into moral panic.

This paper is situated in the continuity of research conducted on Dark Romance and its media representations. It proposes to analyze, using NooJ, the linguistic forms through which the media construct and intensify a register of fear, with particular attention paid to the gradation of the vocabulary mobilized and to the resulting effects of dramatization. The aim here is to describe the linguistic mechanisms through which a cultural object becomes progressively problematized in the media sphere, rather than to evaluate the literary genre as such, which belongs to another strand of research.

The corpus under study consists of approximately thirty different French-language media outlets, observed over several years, including general and specialized print press, online media, audiovisual programs, and podcasts, selected for their explicit treatment of Dark Romance during this recent period. This corpus makes it possible to observe the circulation of a shared lexicon, while also highlighting variations linked to formats, media supports, and enunciative postures.

The analysis is based on an instrumented exploitation of the corpus using NooJ, mobilized as a tool for lexical identification, the structuring of formulations, and the comparison of lexical and discursive units. Particular attention is paid to terms belonging to the semantic fields of fear, danger, influence, and threat, as well as to their syntactic combinations and contexts of enunciation. The operations of extraction and contextualization make it possible to objectify lexical regularities and to trace their progressive intensification in media discourse.

The study thus brings to light a lexical gradation, ranging from formulations associated with relatively ordinary cultural criticism to markedly more

dramatizing expressions, contributing to the establishment of a lasting imaginary of danger. This gradation participates in the construction of an interpretative framework in which Dark Romance is associated with recurrent figures such as youth vulnerability, the supposed influence of literary works, or the fear of a normalization of behaviors deemed problematic.

By showing how these lexical shifts are organized and stabilized within media discourse, this paper proposes a linguistic and discursive reading of the mechanisms of media and moral panic. It highlights the value of a tool-based approach for finely describing the linguistic construction of contemporary cultural fears, particularly when they concern popular and highly mediatized reading practices.

References

[1] Bigey Magali, 2026. « Quand le désir de lire rencontre l'algorithme : image, affects et biais cognitifs dans les prescriptions #BookTok », acte du colloque De Facto, à paraître

[2] Bigey Magali, 2026. « Lire contre les assignations ? Réceptions adolescentes de la Dark Romance à l'ère post-#MeToo », in Anatomie d'une lutte (Dé-)performer le genre dans les médias populaires (titre provisoire) , Université Sorbonne nouvelle et Panthéon Sorbonne, à paraître

[3] Silberztein Max, 2015. La formalisation des langues : l'approche de NooJ. ISTE : Londres, 449 pages.

Modelling of the predicate structure of a sentence for Belarusian and English using NooJ

Mikita Suprunchuk [1], **Yuras Hetsevich**[2], and **Valery Varanovich**[3]
[1] Belarusian State University, Niezalezhnasci ave. 4, Minsk, Belarus
[2] SoftClub – Software Develpment Center LLC, Minsk, Belarus
3 Independent researcher, Minsk, Belarus

The Belarusian linguistic module for the NooJ platform has been under continuous development since 2010. Last year key improvements included expanding the core text corpus to 1.5 million word forms, augmenting the main dictionary with over 200 new tokens, correcting inaccuracies in the morphological tagset, and developing grammars to resolve homonymy in automatic text analysis [2]. This paper continues to improve NooJ approach to Slavic grammar and reports on further advancements made in 2025 and 2026, focusing on enhancing computational semantic and syntactic resources for the Belarusian language.

Given that the sentence is the fundamental unit of syntax, its modeling is central to our work. A sentence differs from a word in several properties: predicativity, the ability to be a unit of communication, etc. A sentence is usually larger than a word, i. e. it has several components that are interconnected. There are a few approaches to describing syntactic connections. Noam Chomsky's generative grammar is the most popular in the USA [1], Lucien Tesnière's dependency theory is widely known in Europe [4], and Igor Melchuk's framework "Meaning ⇔ Text" is frequently applied to Russian, English, and French [3]. Thus, a scheme (or a model) of a sentence is now most often built using the apparatus of direct components or the apparatus of syntactic dependencies. Our analysis of English and Slavic sentence construction utilizes both approaches.

To facilitate this analysis, we have constructed a new multilingual parallel corpus of stylistically diverse texts in Belarusian, English, and Russian. The sentences in this corpus are characterized by high syntactic complexity: Russian source sentences contain 35–75 words, their Belarusian translations 39–70 words, and English translations 56–110 words. The number of predicative parts ranges from 2 to 7 across the languages.

The corpus exhibits frequent instances of inversion (most consistently in English), multiple finite verb forms (up to 7 per sentence). The corpus also regularly presents causal-conditional constructions (up to 10 cases in individual sentences), adverbs and gerunds (in English texts). Temporal coordination of actions (precedence, simultaneity, succession) is also regularly present.

Building upon this corpus, the next phase of our work involves developing computational grammars within NooJ. These algorithms will automatically recognize, quantify, and classify the described syntactic phenomena — such as predicate sequencing, the number and sequence of clauses, inversion, clausal and temporal complexity — while accounting for cross-linguistic variation and genre-specific patterns. This work provides a foundational step toward deeper semantic analysis of texts in Slavic languages.

Keywords: Belarusian Language, English Language, Linguistic Resources, Natural Language Processing, Predicate, Slavic Languages, Sentence Structure, Syntactic Grammar.

References

[1] Noam Chomsky. 2002. Syntactic structures. Mouton, London, UK. 2002.
[2] Yuras Hetsevich, Valery Varanovich, and Mikita Suprunchuk. 2026. Updating the Belarusian NooJ module: integrating new lexical and grammatical resources. In Formalizing natural languages: applications to natural language processing and digital humanities: 19th International conference, NooJ 2025, revised selected papers / eds. by D. Petković, O. Saint-Joanis, M. Silberztein, pages 53–64. https://doi.org/10.1007/978-3-032-17103-0_5.
[3] Igor Mel'čuk. 1988. Dependency syntax: theory and practice. The SUNY Press, Albany, N.Y., USA.
[4] Lucien Tesnière. 1959. Éléments de syntaxe structural. Klincksieck, Paris, France.

A Corrector Prototype for Spanish (Corrector LLCIA_UNR)

Andrea Fernanda Rodrigo, Silvia Susana Reyes, Ana Gigon, Leandro Kintal and **Sandra Mendizaba**
Universidad Nacional de Rosario, Argentina

In the framework of our research project, "Towards a change in language teaching and learning using information technologies", [1] our team made use of the corpus we created to identify salient features of Rioplatense Spanish within a specific community: young university students over eighteen years of age [3]. To this end, a Google form was made available to National University of Rosario (UNR) students, who filled it out with a family or school anecdote, providing their informed consent for the use of the collected data in our research. In turn, this corpus was uploaded to the UNR Academic Data Repository (RDA-UNR) [2], where it is freely available. Our fundamental idea was to explore the language used by young people and to identify their own genuine expressions with a dual aim: on the one hand, to empower these idiomatic expressions in terms of their lexical richness, and on the other, to devise strategies for enhancing these texts while preserving their distinctive Latin American linguistic peculiarities.

In 2025, the creation of the *Laboratorio de Lingüística Computacional e Inteligencia Artificial* (LLCIA) at the Faculty of Humanities and Arts (UNR) prompted us to reconsider our work, deepening and further advancing the initiatives previously undertaken at the *Centro de Estudios sobre Tecnología Educativa y Herramientas Informáticas de Procesamiento del Lenguaje* (CETEHIPL). This rethinking fostered an interdisciplinary approach to our research, integrating the contributions of computer scientists with the advances already achieved by our team of humanists. As a result of this joint effort, we focused on a specific feature of the corpus: the marked relaxation of subject-verb agreement rules, identified in an earlier analysis [3]. Thus, the interdisciplinary team set out to address this deficiency through the development of a specialized corrector (Corrector_LLCIA-UNR), tailored to the characteristics of Argentinian Spanish, that would:

[1] Our four-year research project (2023-2026), hosted by the Faculty of Humanities and Arts, was approved by the Supreme Council Resolution RCS No. 338/2023 of the National University of Rosario (UNR), Argentina.

- Be grounded in Latin American cultural reality in a broad sense, taking into account the idiosyncratic characteristics of the region, with particular emphasis on Argentina.
- Provide the explanations required by the user from a wide-ranging perspective. Recognize the multiple possibilities of text construction. Most current correctors operate under a one-and-only-one model: they offer a single, exclusive correction for each error, with no ambiguity or multiplicity. In contrast, our intention is to reverse this trend. We aim to develop a user-friendly computational tool that provides multiple possible responses simultaneously.
- Encourage the author's creativity by offering educational feedback that enables the user to enhance their linguistic output and overcome shortcomings in a dynamic and straightforward manner without requiring advanced knowledge of grammar.

Our corrector will rely on the dictionaries and grammars developed for the Spanish Module Argentina [1], optimizing Nooj's XML output to collect both syntactic and lexical data. In this presentation, we will share the initial progress of this prototype, as the production process is still in its early stages.

Keywords: Corrector, Rioplatense Spanish, youth grammar.

References

[1] Spanish Module Argentina, https://nooj.univ-fcomte.fr/resources.html.
[2] Rodrigo, Andrea Fernanda, 2025, "Dataset corpus de textos escritos por estudiantes de grado de la Universidad Nacional de Rosa\rio", https://doi.org/10.57715/UNR/5KKC9I, RDA UNR, V1.
[3] Rodrigo, A.F., Reyes, S.S., González, M. (2026). Formalizing the Language of Young University Students. In: Petković, D., Saint-Joanis, O., Silberztein, M. (eds) Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2025. Communications in Computer and Information Science, vol 2832. Springer, Cham. https://doi.org/10.1007/978-3-032-17103-0_9.

The Ball is Round: Idioms in Sports Journalism - SuperSport Croatian Football Cup

Katja Bešlić
Faculty of Humanities and Social Sciences, Zagreb, Croatia

This paper presents the use of established idioms in sports journalism. The analysis of idioms in sports journalism is relevant because it enables the understanding of the Croatian language and journalistic styles, as well as their analysis in other corpora. The project aims to conduct a structural and semantic analysis of the idioms found in the description of a football match, player, team, or event within the game. This corpus and grammar contribute to a better classification and description of idiom forms.
The analysis of sports journalistic style is a widespread topic, written about in professional and entertainment articles, but the topic has not been approached in this way in Croatian so far. The motivation behind choosing this topic lies in the social role of football and the stylistic nature of journalism, which provides a "*fertile ground*" for a high number of styles. Prof. Ivan Marković processes and catalogues the stylemes of sports reports, including idioms, while prof. Ida Raffaelli and Daniela Katunar approach conceptual metaphors in Croatian sports discourse, with an emphasis on the domain of football. This topic has received more attention in other languages, such as English and German, while this work can provide a foundation for further processing of sports journalism. The sample of texts for the corpus was collected via the Internet from online available newspaper and portal articles, with the original texts being transcribed. The period was November and December 2025. These were matches between Croatian clubs such as GNK Dinamo and NK Karlovac 1919 during the Cup quarter-final qualifications. The project corpus does not analyse player and coach comments, but only newspaper reports, and contains over 12,000 tokens. A syntactic grammar with sub-branches was created that automatically recognises and annotates idioms. The grammar contains the structure of idioms, and using the included texts, it is possible to locate and describe them contextually. Thus, idioms are divided into two categories: according to their theme (general, such as "to be on thin ice", and sports-related, such as "to dominate the terrain") and according to their verbal structure (containing either a single verb or multiple verbs). Using the NooJ environment, a large corpus of sports reports from 18 sources was created, containing idioms and a grammar that parses and catalogs idioms by their structure. This enables stylistic comparison of different and

identical sports media and the improvement of journalistic reporting and translation.

Keywords: Phraseology, Football, Sports Reporting, Journalistic Style, NooJ, Croatian.

References

[1] Katunar, D., Raffaelli, I. (2016). A discourse approach to conceptual metaphors: A corpus-based analysis of sports discourse in Croatian. Studia Linguistica Universitatis Iagellonicae Cracoviensis 133 (pp. 136-144).
[2] Marković, I. (2024). Stilovi suvremene hrvatske nogometne kolumne. Hrvatsko filološko društvo (pp. 13-15, 63-71).

Automatic retrieval of Quechua adjectival forms

Maximiliano Duran
Université de Franche-Comté. Besançon, France

Adjectives in Quechua can be derived and inflected using a set of appropriate suffixes. We present the formalization of the morpho-syntactic grammars describing them. This formalization, necessary for the Quechua Language Processing (QLP), had not yet been proposed.
These grammars can generate thousands of derived and inflected forms, and can also serve to retrieve any adjectival form in a written corpus. Concerning the inflection, let's take a qualitative adjective like *sumaj* [beautiful], and apply one of the suffixes: {*cha, lla, yay, yachiy*}, we shall obtain the following inflected forms: *sumaj**cha*** [small beautiful one], *sumaj**lla*** [in a beautiful way], or derived forms *sumaj**yay*** [to become beautiful], *sumaj**yachiy*** [to adorn].
We have manually inventoried all suffixes allowing obtaining inflections and derivations of adjectives. We found 39 suffixes for adjectives ending in vowels and 41 for those ending in a consonant. They are homonyms with the same subset of nominal suffixes; actually, they are semantically different. They play a fundamental role in adjective morphology, and have repercussions, among other things, on a significant increase in the language's lexicon through the generation of new adjectives, adverbs, and derived verbs out of simple POS's.
A notable feature of Quechua is that these suffixes can be applied alone or in combinations of two, three, and even four of them. For example, the following combinations containing two, three, and four suffixes: {***cha-ta, cha-lla-ta, kuna-lla-ta-raj***} will produce the following inflections: *sumaj**chata*** [the beautiful one], *sumaj**challata*** [in a good way], *sumaj**kunallataraj*** [primarily those who are pretty]. In this paper, we show how we have constructed all these valid combinations. Then how did we build their corresponding paradigms. Then, we constructed the general formal grammar to inflect simple adjectives ending in a vowel or in a consonant.
We began by manually constructing a Boolean matrix that describes the valid bi-suffix combinations [1,3]. The tri-suffix, and four-suffix combinations were obtained using simple matrix operations. Then, using NooJ's formalism [5], we constructed the corresponding formalized paradigms and grammars, allowing the generation of inflected and derived forms: A_V and A_C.

If we apply the grammar A_V_1, which includes a single suffix, to the adjective ending in a vowel like *taksa* [small], we obtain 41 inflections.
The grammar A_V_2 that includes paradigms with two suffixes generate 1054 grammatically valid inflections, and the one corresponding to paradigms A_V_3, containing three suffixes yields 7648 inflections.
We have also obtained the formalized grammas for the following derivations:
A>V, an adjective that becomes a verb: *chiri* [cold] → *chiri**chiy*** [to refrigerate].
A>N, an adjective that becomes a noun. In fact, most of the adjectival suffixes will derive the adjective into a noun, as shown in the following example: *taksa**ta*** [the little one](N).
N>A, a noun that becomes an adjective: wayra → *wayra**nirai*** [light as the wind].
V>A, a verb that becomes an adjective: *tukuy* [to finish] → *tuku**sja***[finished].
To test the soundness of our grammars, we have applied them to our corpus, a collection of short stories [4], in order to retrieve all the adjectival forms. We found 226 adjectival forms out of 242 actually present with noun and adverbial forms, this represents a 0,93% success. The non-recognized ones are mainly due to ambiguities, which demand the use of disambiguation grammars.

Keywords: Automatic adjectival retrieval, adjectival suffixes, adjective inflection, adjective derivation.

References

[1] Maximiliano Duran. 2014. Formalizing Quechua verbs Inflexion. Proceedings of the NooJ 2013 International Conference, Saarbrücken.
[2] Maximiliano Duran. 2016. Quechua Module for NooJ Multilingual Linguistic Resources for MT. In: Barone, L., Monteleone, M., Silberztein, M. (eds) Automatic Processing of Natural-Language Electronic Texts with NooJ. NooJ 2016. Communications in Computer and Information Science, vol 667. Springer, Cham. https://doi.org/10.1007/978-3-319-55002-2_5.
[3] Maximiliano Duran, 2017. Dictionnaire électronique français-quechua des verbes pour le TAL. Thèse Doctoral. Université de Franche-Comté. Mars 2017.
[4] Porfirio Meneses, 2001. Sept contes en quechua (1. Yupinta, 2. Hukpa wasin, Kawsayninchik kutimuptin, 4. Muti maskaypi, 5. Tayta Matias, 6. Warmichaykita waylluy, 7. Chiqnipacha. Edition quechua-français. C. Itier. Published by Langues et Mondes. Paris.

[5] Max Silberztein.2016. Language formalization: the NooJ's Approach. Wiley Eds. London.

Warriors and Princesses: Gender Patterns of Metaphors in Croatian Sports Journalism

Viktorija Borko
Faculty of Humanities and Social Sciences, Zagreb, Croatia

This study explores gendered metaphor patterns in Croatian sports journalism, examining how male and female athletes are differently represented through war, animal, aesthetic, and emotional metaphors. Although such topics have been widely studied in Anglo-American discourse analysis, Croatian sports discourse research using corpus tools such as NooJ remains limited. The corpus includes approximately 15,000 words from articles published between 2024 and early 2026 on the portals *Sportske novosti* and *Germanijak.hr*, focusing on major Croatian sporting events and athletes. Using the NooJ tool, a grammar for identifying "adjective + noun" constructions was developed, achieving 93% precision and enabling a more objective analysis of gender representation in sports discourse, in line with the findings of Lidija Ban Matovac [1] on the role of lexical choices in perpetuating traditional gender roles.

The analysis revealed a clear dichotomy in descriptions. In the case of female athletes, metaphors emphasizing appearance, elegance, and emotions dominate. Examples such as *Snow Queen, golden girl, and junior princess idealize success*, while descriptions such as gentle or sweet girl focus on personality rather than physical strength. Even when terminology associated with power is used (*fierce girl*), it is often softened through diminutives or emotional framing. In contrast, male athletic success is framed as a state of war. War metaphors prevail: *battle, uncompromising charge, blue army, heavy artillery*, and *Croatian warrior*. Male athletes are dehumanized through mechanical metaphors (*tireless machine*) or glorified as archetypal predators (*young lion*).

This representation confirms the arguments presented by Cooky [2], suggesting that media reporting systematically normalizes masculinity as aggressive and dominant, while women's sports are treated as aesthetic or emotional events, thereby maintaining hierarchies of power within the sports world. The study confirms that metaphorical language in Croatian media is not neutral, as it systematically reproduces gender patterns in the representation of athletes. The use of NooJ proved effective in identifying these discursive structures and offers strong potential for further large-scale corpus analysis.

Keywords: Croatian, NooJ, discourse analysis, representation of athletes, gender stereotypes

References

[1] L. Ban Matovac. Metaphorical Models in the Language of Sport. Hrvatska sveučilišna naklada, Zagreb, 2012. (This work provides a detailed analysis of how metaphors in the Croatian language shape sporting reality.)
[2] C. Cooky. Women, Sports, and Activism. In The Oxford Handbook of Sport and Society. Oxford University Press, 2018. (The author explores sociological aspects and the media marginalization of female athletes in contemporary society.)

Cantada a bolu, Computational Analysis of Sardinian Extemporaneous Poetry

Silvio Calderaro[1,2] and **Johanna Monti**[1]
[1]Università di Napoli L'Orientale, Naples, Italy
[2] Università di Pisa, Pisa, Italy

This research presents an integrated theoretical-methodological framework for the creation and computational analysis of a dataset of Sardinian extemporaneous poetry, an oral tradition characterized by versified improvisation and rigorous metrical conventions. The study addresses the problem of constructing a structured digital corpus from oral performances, implementing a rule-based linguistic analysis architecture through the NooJ platform [5] to model the multiple levels of complexity of Sardinian [1].

While recent neural approaches have begun to address core NLP tasks for Sardinian, such as part-of-speech tagging [3] and the development of lexical resources [2], the lack of structured corpora for oral poetic performance remains a significant gap. To address this, the core of our research consists of constructing an annotated dataset of improvised octaves and developing targeted computational linguistic resources. In this phase, our methodology leverages NooJ to explicitly formalize the structural constraints that define the Sardinian octave: rhythmic patterns, syllable count, and rhyme schemes.

The analysis of the dataset through NooJ allows us to treat extemporaneous oral discourse as a systematically annotatable object, making explicit the structural logic that governs the cantada. The use of targeted local grammars enables a granular analysis of the interaction between the rigidity of metrical constraints and the creative freedom of improvisers. Furthermore, to investigate the formulaic nature of the compositions, the rule-based annotations are combined with the extraction of n-grams and the statistical analysis of word occurrences. This approach facilitates the automatic identification and quantification of recurrent multi-word blocks and rhetorical strategies that characterize the genre through systematic corpus interrogation.

Beyond structural analysis, the contribution explores the pragmatic-semantic dynamics of poetic dueling. By integrating specialized electronic lexicons within the NooJ environment, we examine how specific lexical choices operate within the agonistic context of the performance. The results highlight the effectiveness of the NooJ-based approach in the study of low-resource languages, where the scarcity of digitalized resources limits the applicability of purely statistical data-driven approaches [4]. The

creation of a computationally annotated dataset and its analysis through formal linguistic resources provide a rigorous representation of the structural and linguistic architecture of this tradition, proposing a replicable model for the documentation and scientific analysis of complex oral traditions at risk of extinction.

References

[1] Eduardo Blasco Ferrer, Peter Koch, and Daniela Marzo, editors. Manuale di linguistica sarda. De Gruyter, Berlin/Boston, 2021.

[2] Marco Angioni, Francesco Tuveri, Maurizio Virdis, Luca L. Lai, and Marco E. Maltesi. Sardanet: a linguistic resource for sardinian language. In Proceedings of the 9th Global Wordnet Conference (GWC 2018), 2018.

[3] Simone M. Carta, Fabrizio Concas, Gianni Fenu, Andrea Giuliani, Marco M. Manca, Mirko Marras, and Simone Pisano. A BERT-based approach for part-of-speech tagging in the low-resource context of sardinian. In Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[4] Barbara Plank. What to do about non-standard (or non-canonical) language in NLP. In Proceedings of KONVENS 2016, Bochum, Germany, 2016.

[5] Max Silberztein. Formalizing Natural Languages: The NooJ Approach. ISTE Press and John Wiley & Sons, London, 2016.

Noun phrase complexity in a corpus of children's writing

Daša Farkaš and **Kristina Kocijan**
University of Zagreb, Faculty of Humanities and Social Sciences,
Department of Linguistics, Zagreb, Croatia

Syntactic development is a long-term process in which linguistic constructions become increasingly complex as children mature, progressing from the acquisition of basic structures to their integration into more hierarchically organized units. Although numerous studies have been conducted, there is still no consensus on which measures of syntactic complexity most reliably capture developmental progress, particularly when comparing spoken and written modalities [1, 2, 3]. Previous research on Croatian child language relies primarily on spoken corpora within the CHILDES framework, including 1) *Croatian corpus of child language* [4], consisting of spontaneous interactions of three children, which was later completed with new records [5]; 2) *Croatian Corpus of Preschool Child Language* [6], containing transcripts from 60 monolingual speakers of Croatian; and 3) *Croatian Frog Story Corpus* [7], comprising 157 narrative transcripts of monolingual speakers of Croatian, aged 3 to 11. These resources offer insights into early syntactic development but focus on spoken language and clause-level measures [3, 8]. However, they do not capture the developmental trajectory of written syntax, which becomes increasingly relevant as children progress through formal schooling.

This study addresses that gap by examining noun phrase (NP) complexity in children's writing, focusing on developmental differences between pupils in the 5th and 8th grades of Croatian primary schools. The analysis is based on the *ŠKOLARAC <traces in words>* corpus, a newly developed computational resource containing systematically collected and processed written assignments produced by pupils across multiple grade levels. Unlike previous Croatian corpora, ŠKOLARAC enables large-scale, quantitative investigation of written language and provides a foundation for exploring syntactic development beyond traditional clause-based metrics.

Findings from English [9] show that NP complexity increases across schooling, with older students producing longer and structurally richer noun phrases. Based on these findings, we hypothesize that Croatian pupils exhibit similar developmental patterns and that the transition from early to mid-adolescence is reflected in measurable changes in NP structure.

We analysed NP length, modifier types, and internal structural depth in two ŠKOLARAC subcorpora (5th and 8th grade). To systematically identify and annotate noun phrases and their internal structure, we developed a set of

syntactic grammars in NooJ. These grammars enabled the automatic recognition of base NPs together with their modifiers, allowing for consistent extraction across the corpus. The annotated output was then exported and further processed using a data-visualization tool, which facilitated quantitative comparison of NP features across grade levels.
The results reveal clear developmental differences between the age groups: NP complexity increases with age and also shifts from modifying the noun with an adjective to expanding the phrase with a PP. This study provides the first corpus-based analysis of NP complexity in Croatian children's writing and demonstrates the value of written corpora for understanding syntactic development. By focusing on noun phrases rather than clause-level measures, the research highlights an aspect of syntactic growth that is particularly relevant for academic writing, literacy development, and computational modelling of child language. The findings contribute to both developmental linguistics and NLP by offering new empirical evidence, a novel corpus resource, and a methodological framework for future studies on written language acquisition.

Keywords: Croatian language, syntactic development, noun phrase complexity, children's writing.

References

[1] Marilyn Nippold. 2006. Language development in School-Age Children, Adolescents, and Adults. Encyclopedia of Language and Linguistics, 368-373.
[2] Ruth A. Berman. 2018. Development of complex syntax: From early clause-combining to text-embedded syntactic packaging. Handbook of Communication Disorder, ed. Bar-On Amalia and Dorit Ravid, Berlin, Boston, 235-256.
[3] Magdalena Bedeković, Gordana Hržica, and Matea Kramarić. 2021. Analiza sintaktičke složenosti dječjeg pripovjednog diskursa. FLUMINENSIA, 33 (2), 417-443. https://doi.org/10.31820/f.33.2.8
[4] Melita Kovačević. 2002. Croatian corpus, CHILDES. https://talkbank.org/childes/access/Slavic/Croatian/Kovacevic.html
[5] Antonia Ordulj, and Gordana Hržica. 2015. Obnavljanje Hrvatskog korpusa dječjega jezika. Logopedija, 5 (1), 25-31. https://hrcak.srce.hr/140212
[6] Gordana Hržica, Sara Košutar, Tomislava Bošnjak Botica, and Petar Milin. 2024. The role of entrenchment and schematisation in the acquisition

of rich verbal morphology. Cognitive Linguistics. https://doi.org/10.1515/cog-2023-0022
[7] Ivana Trtanj, and Jelena Kuvač Kraljević. 2017. Language and speech characteristics of children's narratives: the analysis of microstructure. Govor: časopis za fonetiku, 34, 1, 53-69 doi:10.22210/govor.2017.34.03
[8] Jelena Kuvač Kraljević, and Marijan Palmović 2007. Metodologija istraživanja dječjega jezika. Jastrebarsko: Naklada Slap.
[9] Philip Durrant, and Mark Brenchley 2023. Development of Noun Phrase Complexity Across Genres in Children's Writing. Applied Linguistics, Volume 44, Issue 2, 239–264.

Topological Braiding of Linguistic Structures: Anyons, Recursion and Symbolic Robustness in NooJ

Ritamaria Bucciarelli[1], **Francesco Terrone**[2], **Colomba La Ragione**[3], **Andrea Fernanda Rodrigo**[4], **Javier Julian Enriquez**[5]
[1]Università Ca Foscari di Venezia, Venezia, Italy
[2]Fondazione Francesco Terrone di Ripacandida e Ginestra, Italy
[3]Università Pegaso, Italy
[4]Universidad Nacional de Rosario, Rosario, Argentina
[5]Universitat Politècnica de València, Valencia, Spain

This paper introduces a new approach in computational linguistics called: The Topology of the Word. In this approach, poetic verse and complex sentences are turned into synthetic, fixed, and uniform formal language. Inspired by Michel Planat's theory of non-abelian anyons and graphic coverings, we propose NooJ as a "topological laboratory" capable of validating semantic stability through rigid geometric constraints.
Within this framework, applied through Digital Humanities to the works of Dante Alighieri, linguistic units are identified as points (formal operators) whose interactions generate lines (contexts). Following Planat's exhortations, we define a "line" as a context of at least three points sharing a common property—such as syntactic agreement—where the switching of operators [a, b] = $b^{-1}a^{-1}$ba determines structural coherence. determines structural coherence.
By identifying linguistic mechanisms and inventing a rhetorical methodology designed as an architecture of constraints, we demonstrate how NooJ v7 maps:

1. **Quantum Computation and Fusion Rules**: The detection of data for categoric reduction follows SU (2)K anyon fusion rules (e.g., 1/2$\hat{\otimes}$1/2=0$\hat{\otimes}$1) transforming Dante's syntactic recursion into stable anyonic dynamics.
2. **Recursion and Braiding**: Recursive phenomena and rhetorical figures are interpreted as **braiding processes** that preserve the topological invariant of meaning. If the braid undergoes decoherence, NooJ detects a "topological defect," rigorously identifying semantic hallucinations.
3. **NooJ as a Symbolic Interface**: NooJ acts as an intermediate framework that mediates between complex expressive structures and computable models. It facilitates a selective, theory-driven reduction of structures where rhetoric functions as the generative principle that stabilizes features for advanced structural computation.

4. **Transformation into Formal Language**: Through **Graphic Coverings**, NooJ lifts the linear text string into a multi-sheeted representation, where the verse becomes a homologous mathematical object validated by braid matrices

$$\sigma_1 \text{ and } \sigma_2.$$

This approach, uniting Gödel's logical legacy with anyonic topology, positions NooJ as the premier experimental space for a **Conscious AGI**, where linguistic truth is anchored in the stability of topological form and the precision of symbolic-quantum counting and validation.

Keywords: NooJ v7, Dante Alighieri, Topology of the Word, Rhetoric, Anyons, Braid Groups, Formal Representation, Digital Humanities, Explainable NLP.

References

[1] Bucciarelli, R. (2025). *Recursive Linguistic Structures and Symbolic Computation in AI*. Archivio della Ricerca Ca' Foscari University of Venice (ARCA).
[2] Bucciarelli, R. (2025). *Quantum Computation and Measurements from a Space-Time in Fixed Languages*. ResearchGate.
[3] Bucciarelli, R. (2025). *Topologies of the Mind. Language, Anyons, and Symbolic Consciousness*. HAL Open Science.
[4] Chomsky, N. (2014). *Aspects of the Theory of Syntax*. MIT press.
[5] Gödel, K. (1992). *On formally undecidable propositions of Principia Mathematica and related systems*. Courier Corporation.
[6] Goodman, J., Vlachos, A., & Naradowsky, J. (2016, August). Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1-11).
[7] Gusfield, D. (1997). Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News*, *28*(4), 41-60.
[8] Hofstadter, D. R. (1999). *Gödel, Escher, Bach: an eternal golden braid*. Basic books.
[9] Jurafsky, D., Martin, J.H. (2023). *Speech and Language Processing*. 3rd edn. Draft version.
[10] Minsky, M. 1986. *The Society of Mind*. Simon & Schuster.

[11] Pearl, J. (2018). *The book of why: The new science of cause and effect*. Basic Books.
[12] Planat, M., & Amaral, M. (2024). What ChatGPT has to say about its topological structure: the anyon hypothesis. *Machine Learning and Knowledge Extraction*, *6*(4), 2876-2891.
[13] Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2015). *A new companion to digital humanities*. John Wiley & Sons.
[14] Silberztein, M. (2016). *Formalizing natural languages: The NooJ approach*. John Wiley & Sons.
[15] Terrone, F., Bucciarelli, R., Rodrigo, A.F., Julian Enriquez, J. (2025). *Gödel's Legacy: Formal Thinking, LLM and the Evolution of Artificial Intelligence*. HAL Open Science.
[16] Turing, A. M. (1987). Computing machinery and intelligence (1950). *Mind*, *59*(236), 33-60.

Teaching Latin Morphology through Resource Building in NooJ

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

Latin pedagogy differs significantly from modern foreign language teaching due to the absence of native speakers and the fundamentally diachronic nature of its communication. In recent decades, teaching approaches to Latin have shifted from grammar-translation methods toward strategies that prioritize textual comprehension, active student engagement, and functional mastery of the language system, particularly through reading and communicative approaches [5: 35]. These principles are evident in Croatia's national Latin curricula for both primary and secondary education. However, Latin continues to be perceived as an excessively demanding subject, often associated with rote memorization of paradigms. This perception is particularly pronounced in contexts where Latin is taught as an auxiliary subject, for example, within Romance language programs, where the focus is placed on linguistic structure as the basis of historical grammar, while broader linguistic and cultural dimensions are neglected, further reinforcing the grammatisation of instruction. In contemporary foreign language pedagogy, increasing attention is being paid to the development of teaching competences related to the use of digital tools. Research on the NooJ platform [4] has primarily addressed modern languages (cf. [2]; [1]). For Classical Latin, NooJ lacks comprehensive morphological resources beyond those available for Medieval Latin (cf. [3]: 26), which restricts its direct application to automatic linguistic analysis. Nevertheless, this limitation creates opportunities for active, inquiry-based learning.

This paper presents a teaching model for higher education in which students independently construct limited linguistic resources in NooJ, such as lexical entries and basic morphological graphs, for selected Latin grammatical topics. This process requires students to demonstrate precise knowledge of morphological categories, formalize grammatical rules, and empirically test these rules on authentic Latin texts. Consequently, students shift from passive recipients to active researchers and constructors of a linguistic system. The model results in improved understanding of Latin morphology, enhanced metalinguistic awareness, and increased student engagement compared to traditional paradigm-based instruction.

Keywords: Latin language pedagogy, inquiry-based learning, digital tools in language teaching, NooJ.

References

[1] Andrea Rodrigo and Rodolfo Bonino. 2019. *Aprendo con NooJ. De la lingüística computacional a la enseñanza de la lengua*. Ciudad Gótica, Rosario.

[2] Julia Frigière and Sandrine Fuentes. 2015. Pedagogical Use of NooJ dealing with French as a Foreign Language. In *Formalising Natural Languages with Nooj 2014. Selected papers from the NooJ 2014 International Conference.* Johanna Monti et all. (eds.), pp. 186–197. Cambridge Scholars Publishing, Cambridge.

[3] Linda Mijić and Anita Bartulović. 2024. Recognizing Verbs in Medieval Latin. In *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2024, Revised Selected Papers*, pp. 16–27. Springer, Cham.

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

[5] Steven Hunt. 2016. *Starting to Teach Latin*. Bloomsbury Academic, London.

Improving Croatian NLP Infrastructure: An Expanded and Structurally Annotated Verb Dictionary for NooJ

Krešimir Šojat[1] and **Kristina Kocijan**[2]
[1] University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics, Zagreb, Croatia
[2] University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences, Zagreb, Croatia

The current Croatian verb dictionary available within the NooJ linguistic platform comprises 5,263 lemmas and generates 472,205 inflected forms, offering a solid but incomplete foundation for computational processing of Croatian. Given the morphological richness of Croatian as a Slavic language, characterized by extensive verbal inflection, aspectual distinctions, derivational productivity, and complex interactions between morphology and syntax, substantial lexical coverage and detailed annotation are essential for high-quality natural language processing. This paper presents a major expansion of the Croatian verb dictionary for NooJ [6], adding approximately 10,000 new verb lemmas enriched with structured morphological and syntactic information. This enhancement significantly increases the descriptive adequacy of Croatian resources for NooJ and provides a more robust foundation for linguistic research, corpus annotation, and language technology development.

Each new lemma is annotated with part-of-speech information and several features crucial for modeling Croatian verbal morphology: aspect (perfective, imperfective, biaspectual), reflexivitiy, the corresponding inflectional paradigm [5, 1, 3], and the lexical root [4]. For example, the entry dopisivati is encoded as *dopisivati,V+**Root**=pis+**Aspect**=ipf+**Prelaz**=pov+**FLX**=NARUČIVATI*, enabling NooJ to generate all relevant inflected forms while simultaneously providing structured information for downstream syntactic and semantic processing. The explicit marking of verbal aspect and roots is particularly important in Slavic languages, where aspectual pairing realized through various derivational processes form tightly interwoven morphological systems. Derivational processes between aspectually related verbs comprise prefixation, suffixation, and ablaut, as well as various combinations of these processes.

The methodology for constructing this expanded dictionary combines automatic extraction of verbal lemmas from publicly available monolingual dictionaries and corpora, including the ŠKOLARAC (large Croatian L1

learner corpus), manual analysis and classification, as well as systematic validation using existing Croatian linguistic resources. Special attention is devoted to identifying productive derivational families and normalizing root representations [4], which facilitates the modeling of Croatian verb formation patterns and supports future work on automatic detection of aspectual pairs and derivational networks.

Inflectional paradigms are assigned using a transparent set of criteria grounded in established descriptions of Croatian verbal morphology [2], ensuring consistency and reproducibility. The enriched verb resource improves lemmatization, morphological tagging, and syntactic annotation, enabling more accurate analyses of language, more reliable frequency data, and more nuanced pedagogical applications. The dictionary thus serves not only as a linguistic resource but also as an infrastructural enhancement for corpus-based research and language-learning technologies.

The contribution of this work is threefold. First, it delivers the most extensive and systematically annotated Croatian verb dictionary available for NooJ, increasing the number of lemmas by nearly 200%. Second, it introduces structured morphological and syntactic features that enhance rule-based processing and enable advanced linguistic analyses of the Croatian verbal system. Third, it lays the groundwork for future research, including automatic detection of derivational families, improved semantic tagging, and integration with machine-learning approaches that require high-quality linguistic input. Overall, these improvements significantly strengthen the infrastructure for Croatian language processing within NooJ and support both theoretical linguistic research and practical NLP applications. Furthermore, the structured data provide an essential foundation for developing a rule-based tool for automatic morphological analysis of Croatian verbs within the NooJ platform. Finally, the objective of this study is to present the initial steps toward implementing automatic morphological segmentation of Croatian verbs, based on data analysis aligned with the principles outlined above.

Keywords: verbs, morphology, dictionary, Croatian, NooJ.

References

[1] Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević and Marija Znika. 2005. *Hrvatska gramatika*. Školska knjiga, Zagreb.

[2] Ivan Marković. 2012. *Uvod u jezičnu morfologiju*. Disput, Zagreb.

[3] Josip Silić and Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika - Za gimnazije i visoka učilišta*. Školska knjiga, Zagreb.

[4] Krešimir Šojat, Matea Srebačić and Vanja Štefanec. 2013. CroDeriV and the Morphological Analysis of Croatian verb. *Suvremena lingvistika*, Vol. 39 No. 75

[5] Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian Morphological Lexicon. In: *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*. Association for Computational Linguistics, pages 41–45.

[6] Kristina Vučković, Marko Tadić, and Božo Bekavac. 2010. Croatian Language Resources for NooJ. *CIT - Journal of computing and information technology*, 18: 295-301. doi: 10.2498/cit.1001914

Adapting the Arabic NooJ Module for Persian: A Pilot Study

Farideh Nikpour
Shiraz University, Shiraz, Iran

Persian is a morphologically rich and partially inflectional language, with orthographic and morphological features that differ from, yet partially resemble, Arabic. Despite these similarities, Persian exhibits unique inflectional patterns, compound verbs, and clitic constructions that present challenges for computational processing. NooJ currently offers linguistic modules for multiple languages, including Arabic, French, and English, but no official module exists for Persian. This study presents a pilot effort to adapt the Arabic NooJ module for Persian, establishing a foundation for a dedicated Persian linguistic resource within the NooJ environment.

The main objectives of this study are twofold: (1) to analyze linguistic similarities and differences between Arabic and Persian that impact morphological and syntactic processing, and (2) to adapt the Arabic module by modifying its lexicon, morphological graphs, and syntactic rules for Persian. The adaptation process involved three key stages. First, the Arabic morphological and syntactic graphs were examined to identify reusable structures, including prefix and suffix patterns, tokenization rules, and encoding conventions. Second, a preliminary Persian lexicon was developed, containing approximately 500 frequently used nouns, verbs, and adjectives, each annotated with morphological tags compatible with NooJ standards. Third, a set of initial morphological and syntactic graphs was created to capture key Persian inflectional phenomena, such as plural formation with the suffix "-ها", present tense conjugations with "می –", and past participle constructions.

The adapted module was tested on a small Persian corpus drawn from news articles and short literary texts. Preliminary evaluation shows that roughly 78% of tokens were correctly identified and annotated, demonstrating that Arabic-based infrastructure can be effectively repurposed for Persian with moderate modifications. Challenges were observed in processing compound verbs, handling enclitic pronouns, and addressing orthographic variations (e.g."می‌رود" vs." میرود") indicating areas for future improvement.

This pilot study highlights that NooJ's architecture is flexible and language-independent, capable of supporting Persian linguistic analysis. By leveraging the existing Arabic module as a scaffold, we provide a methodological pathway for extending NooJ to Persian, including the future

development of a comprehensive morphological-syntactic parser and semantic annotation layer. The research contributes a preliminary Persian resource and offers insights into practical strategies for adapting NooJ modules to typologically similar languages, bridging a gap in computational tools for Persian NLP and digital humanities applications.

Naming the Hero: A Computational Study of Referential Variation of Achilles and Odysseus in Italian Homeric Translation

Giulia Speranza
UNIOR NLP Research Group
Università di Napoli "L'Orientale", Naples, Italy

Homeric epic is characterized by extensive referential variation, whereby characters are designated not only by proper names but also by formulaic epithets and by periphrastic expressions encoding genealogical relations and patronymics. Identifying the patterns of epithetic and genealogical reference is essential for capturing the narrative and semantic construction of heroism in the Iliad and the Odyssey. From a computational perspective, epithets and periphrastic expressions involve predictable yet structurally complex patterns, posing challenges for automatic detection, annotation, and quantitative analysis. This study focuses on the formulaic and periphrastic apparatus defining Achilles and Odysseus, as they represent the main heroes in the Iliad and the Odyssey, with the aim of quantifying and classifying the referential strategies that construct heroic identity in Italian translations of Homer, bridging literary interpretation and computational analysis.

The paper addresses three main research questions i) Which classes of referential expressions (proper names, epithets, patronymics, and role-based periphrases) are most frequent for Achilles and Odysseus, and what semantic attributes (physical, cognitive, moral, relational) do these expressions encode? ii) Does the locus of enunciation (narrative discourse vs. direct speech) correlate with systematic differences in the selection of referential expressions for Achilles and Odysseus? iii) Does the distribution of referential expressions differ between the Iliad and the Odyssey for the two heroes? The methodology relies on a dictionary integrated into the NooJ linguistic environment, which includes the proper names as well as a set of mythological entities relevant to genealogical and relational constructions. These entries are annotated with semantic features (e.g., +HERO, +HUMAN, +GOD, +KING). Local grammars were developed to identify epithetic constructions in which a proper name is combined with one or more qualifying adjectives capturing both canonical forms (e.g., il divino Achille, the godlike Achilles) and more extended variants, as well as the type of adjectives modifying the noun (semantically evaluative, physical, moral). A second set of local grammars targets genealogical and

relational expressions (e.g., figlio di Peleo, (“son of Peleus”), and includes variants with optional adjectives (e.g. il valoroso figlio di Peleo, “the great son of Peleus”) as well as syntactic re-orderings resulting from anastrophe and hyperbaton (e.g., *di Laerte figlio, “of Laertes son”*), and relations with cities (e.g., *Itaca, “Ithaca”*) or populations (e.g., *Mirmidoni, “Myrmidons”*). A subset of the manually annotated corpus was used to validate the automatic extraction, achieving a precision of 88% and a recall of 82% for referential patterns. Quantitative and qualitative analyses of the results show that Achilles has the largest number of distinct referential variants and the highest type/token ratio of epithetic forms, realized across a wide array of grammatical patterns. Epithet-adjective constructions found in the Iliad primarily encode evaluative and physical attributes, emphasizing his valor and martial prowess, while genealogical periphrases are more frequent in the Odyssey, highlighting his heroic lineage. By contrast, Odysseus is predominantly associated with epithets emphasizing cognitive, evaluative, and moral attributes, reflecting his intelligence and strategic acumen. Genealogical constructions are comparatively rare, consistent with the narrative focus on personal agency rather than lineage. In both cases, epithets predominate in narrative passages, whereas in dialogue or direct speech, proper names are used more frequently, illustrating the role of discourse context in Homeric referential choice for the two heroes.

Keywords: Referential Variation, Epithets, Linguistic Patterns, Automatic Extraction

References

[1] J. Aoun and L. Choueiri. Epithets. Natural Language & Linguistic Theory, 18(1):1–39, 2000.
[2] N. Austin. The epithets in homer: A study in poetic values, 1985.
[3] J. H. Dee. The Epithetic Phrases for the Homeric Gods: A Repertory of the Descriptive Expressions of the Divinities of the Iliad and the Odyssey. Taylor & Francis, 2020.
[4] M. Finkelberg. Formulaic and nonformulaic elements in homer. Classical Philology, 84(3):179–197, 1989.
[5] R. H. Gibson. Does hector’s helmet flash? the fate of the fixed epithet in the modern english homer. Oral Tradition, 33(1):89–114, 2019.
[6] I. S. F. Grey. Communicating identity: the significance of epithets in Homer’s Odyssey. PhD thesis, University of Birmingham, 2020.
[7] Z. T. Koblanov, U. Khudaybergenova, R. Kenzhebayeva, P. S. Oteniyazovna, T. Z. Baymagambetovna, and Z. Kozhikbaeva. The

language of heroism: Linguistic and artistic dimensions in homer's iliad and odyssey. Jurnal Arbitrer, 12(4):582–596, 2025.
[8] K. Malone. Epithet and eponym. Names, 2(2):109–112, 1954.
[9] R. Mitchell-Boyask. Odysseus laertiades: Gardening, paternity and the namings of odysseus. Classical World, 118(2):157–182, 2025.
[10] J. V. O'Brien. The transformation of Hera: A study of ritual, hero, and the goddess in the Iliad. Bloomsbury Publishing PLC, 1993.
[11] M. Parry. The homeric gloss: A study in word-sense. Transactions and Proceedings of the American Philological Association, 59:233–247, 1928.
[12] M. Parry and A. Parry. The making of Homeric verse: The collected papers of Milman Parry. Oxford University Press, 1987.
[13] W. M. Sale. The formularity of the place phrases of the iliad. Transactions of the American Philological Association (1974-), 117:21–50, 1987.
[14] S. Sarwar, I. Aziz, R. Kanwal, and M. A. Lodhi. Digital stylistic insights into homer's iliad: A lexical, semantic, and contextual analysis. Journal of Applied Linguistics and TESOL (JALT), 8(4):81–98, 2025.

Say It Again: A Semantic Rule-Based Taxonomy of Medical Term Simplification in Italian Patient Leaflets

Maria Pia di Buono[1], Mariapia Battipaglia[1,2]
[1]Università di Napoli "L'Orientale", Naples, Italy
[2]Università di Pisa, Pisa, Italy

Medical language is a highly specialized discourse characterized by the use of jargon and complex lexicon. As pointed out by several scholars focusing on Italian, e.g., [1,10,4], the lexicon of this language for specific purposes features neoclassical formative elements [8], eponyms and collateral technicalisms. Beyond the lexical level, its complexity also depends on lengthy sentences, relational adjectives and passive forms which increase textual opacity. Such linguistic characteristics create significant barriers for patients, who often find patient-facing texts dense and difficult to process [7].

Despite the considerable efforts undertaken to promote the adoption of plain language principles — including the activities of the International Plain Language Federation and the development of a dedicated ISO standard — patient-oriented communication still requires improvement across its four pillars: relevance, findability, understandability, and usability.

Several studies on medical text simplification have addressed the development of linguistic resources, the identification of medical jargon, and the generation of paraphrases (among others, [5,2,6]).

In the Italian context, di Buono [3] introduced Italian Medical Term Simplification (I-MTS), the first resource specifically designed for medical term simplification. Derived from existing patient information leaflets (PILs), I-MTS contains 1,356 term-simplification pairs organized into three semantic classes: Disease, Medication and Body Part.

Building upon I-MTS, this study investigates class-specific linguistic patterns of simplifications and the semantic relations they encode, as different classes present distinct linguistic behaviours and levels of granularity.

Disease simplifications frequently encode relations such as Affect, Cause, and Caused-by, as in *infezione dei polmoni causata da germi* ("lung infection caused by germs") for the term *polmonite batterica* ("bacterial pneumonia").

Medication simplifications commonly instantiate relations including Belongs-to-a-Class, Cures, Affect, and Causes, often realized through hyperonymy and purpose relations, e.g., *medicinali antinfiammatori*

("anti-inflammatory medicines") or *medicinali utilizzati per trattare le infiammazioni* ("medicines used to treat inflammations").
Body Part simplifications predominantly express relations such as Is-A, Has-Location, and Has-Function, e.g., the descriptive simplification *osso lungo della gamba* ("long bone of the leg") realizes an Is-A relation with *femore*, or *ghiandola coinvolta nella digestione dei cibi* ("gland involved in the digestion of food") instantiates a Has-Function relation with the term *pancreas*.
Starting from the term-simplification pairs attested in I-MTS, we analyze reformulation structures across the three semantic classes and identify recurrent semantic and linguistic patterns associated with each category.
Such patterns constitute the basis for a rule-based framework implemented in NooJ [9] to develop a hierarchy of local grammars organized according to the I-MTS classes. Each grammar is designed to identify a specific semantic relation through class-specific selectional restrictions and to produce the corresponding annotation.
The resulting output enriches I-MTS with a new semantic annotation layer linking terms, semantic relations, and linguistic patterns. Beyond extending the descriptive coverage of the resource, the proposed framework provides a reusable methodology for capturing simplification phenomena in patient-oriented medical communication and paves the way for future NooJ-based applications supporting syntactic and semantic text simplification.

Keywords: Term Simplification, Semantic Relations, Annotation, Linguistic Resources

References

[1] Cortelazzo, M. (1994). Lingue speciali. La dimensione verticale. Padova: Unipress.
[2] Devaraj, A., Wallace, B. C., Marshall, I. J., & Li, J. J. (2021). Paragraph-level Simplification of: Medical Texts. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4972–4984,Stroudsbourg: Association for Computational Linguistics.
[3] Di Buono, M. P. (2026). Italian Medical Term Simplification: From Patient Information Leaflets to Simplified Language Resources. In Proceedings of CL4HEALTH Workshop @ LREC 2026, pages 182-189.
[4] Gualdo, R., Telve, S. (2021). Linguaggi specialistici dell'italiano. Roma: Carocci.

[5] Kandula, S., Curtis, D., & Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2010, 366–370. Bethesda: American Medical Informatics Association.
[6] Kwon S., Yao Z., Jordan H. S., Levy D. A., Corner B., and Yu H. 2022. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. In Proceedings of the Conference on Empirical Methods in NaturalLanguage Processing. Conference on EmpiricalMethods in Natural Language Processing, volume 2022, page 11733.
[7] March-Cerdá JC., Prieto-Rodríguez MA., Ruiz-Azarola A., Simón-Lorda P., Barrio-Cantalejo I., Danet A. (2010). Mejora de la información sanitaria contenida en los prospectos de los medicamentos: expectativas de pacientes y de profesionales sanitarios. Atención Primaria, 42, (1) 22-27. Barcelona: Elsevier España.
[8] Rainer, F., Thornton, A. M. (2002). Morfologia. Linguistica italiana alle soglie del 2000: 1987-1997 e oltre. Pubblicazioni della Società linguistica italiana (44), 1000-1047. Roma: Bulzoni.
[9] Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach. Hoboken: Wiley.
[10] Serianni, L. (2012). Italiani scritti. Bologna: il Mulino.

Grammatical Words and Stop Words

Max Silberztein
Université de Franche-Comté, Besançon, France

Natural Language Processing software applications must distinguish between terms that convey relevant semantic information and words that convey no meaning. For example, automatic summarizers typically filter out empty words, and machine translation software typically does not translate grammatical words.

To process these empty words, most NLP applications use reference lists that contain “stop words”. For instance, users of the Python NLTK package typically start processing a text by performing a “text normalization” step that includes filtering out all the word forms that are included in NLTK’s list of stop words (200 for English), e.g.: a, about, above, after, again, against, ain, all, am, an, and, any, are, etc.

We will show that this approach is inadequate for three reasons:

- These lists of stop words do not include numerous grammatical words, such as adverbs, determiners, particles, prepositions, pronouns and verbs, although these words have long been listed in available open-source electronic dictionaries (Courtois, Silberztein 1990). For instance, NLTK list of English stop words does not include aback (Adverb), another (Determiner), ahead (Particle), anybody (Pronoun), across (Preposition), may (Verb).
- Conversely, many words in this list often appear in multiword units that are meaningful and should not be filtered out. For example, the word have is usually an auxiliary verb, but can also be meaningful in some contexts:
  I have already eaten → have is an auxiliary verb I have a car → have is a meaningful verb.
- Many so-called stop words are homographs and may carry some meaning in specialty languages, e.g., “from” and “to” in a text dedicated to selling train tickets.

We will show that, to filter out meaningless words from a text, it is necessary to perform linguistic analysis, and we will present several resources (dictionaries and grammars) that can be used for this purpose.

References

Courtois, B. & Silberztein, M. Eds.,1990. Dictionnaires électroniques du français. Langue française #87.

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

Marzieh RABIEI
Université Bourgogne Franche-Comté, Besançon, France

This paper presents the computational modeling of Persian nouns in NooJ, as the second stage of a broader project on the formalization of Modern Persian. Following the formalization of the Persian verbal system, this study extends the existing morphological framework to another major lexical category of the language.

Persian nominal morphology raises specific challenges, including plural formation, possessive suffixes, the ezâfe construction, and ambiguity caused by the absence of short vowels in standard writing. These phenomena make automatic morphological and syntactic analysis particularly complex.

The study adopts the same linguistically driven methodology previously applied to the verbal system. A dedicated nominal dictionary is being constructed, together with inflectional and morphological grammars covering plural markers, possessive constructions, nominal derivation, and their interaction with ezâfe. Particular attention is paid to the compatibility of these nominal resources with the existing verbal resources developed in NooJ.

This rule-based approach relies on explicit paradigms and reusable grammars rather than statistical methods. It is therefore consistent with the NooJ framework and particularly suitable for Persian, an under-resourced language in NLP. Preliminary experiments indicate that the integration of nominal resources improves morphological analysis and paves the way for richer syntactic annotation.

By extending the formalization from verbs to nouns, this study contributes to the construction of a coherent and extensible set of linguistic resources for Modern Persian.

The resulting resources support future applications such as parsing, information extraction, and ambiguity resolution, while demonstrating the relevance of NooJ for the systematic modeling of complex linguistic phenomena.

Keywords: Persian language, NooJ platform, nominal morphology, computational linguistics, Persian nouns, rule-based linguistic resources.

References

[1] Gilbert Lazard. 1957. Grammaire du persan contemporain. Librairie C. Klincksieck, Paris.

[2] Pollet Samvelian. In press. Specific features of Persian syntax: The ezâfe construction, differential object marking and complex predicates. In Anousha Sedighi and Pouneh Shabani-Jadidi, editors, The Oxford Handbook of Persian Linguistics, pages 226–269. Oxford University Press.

[3] Max Silberztein. 2016. Formalizing Natural Languages: The NooJ Approach. Wiley, Hoboken, NJ.

A Computational Approach to Optimize Verbal Inflectional Resources with NooJ

Agnieszka K. Kaliska
Adam Mickiewicz University
ul. Wieniawskiego 1, 61712 Poznań

The first Polish resources designed for use in the NooJ system were developed by K.Bogacki (2008), who continued the work with E. Gwiazdecka in the following years [1, 2]. The development of these resources—including a dictionary and a morphological analyzer—was based on excerpts from two well-known novels: Faraon by H. Sienkiewicz (1846-1916) and Sklepy cynamonowe by B. Schulz (1892-1942). This initial corpus, comprising approximately 18k words, resulted in a first NooJ dictionary containing around 1,391 entries, including 280 verbal entries and a corresponding set of 118 inflectional paradigms (see NooJ Linguistic Resources [hereafter: NLR]; access: 07-05-2026).
This study aims to develop the existing NooJ model of Polish verbal inflection.
The motivation for developing this component of the resource stemmed from the intention to exploit the existing NooJ resources for the extraction of motion-related information, specifically manners of motion. From a sample of 848 motion verbs compiled on the basis of open resources from the plWordNet [4], only 37 forms were successfully recognized by the morphological analyzer NooJ for Polish (see NLR). This outcome revealed substantial gaps in the existing dictionary, and highlighted the need to extend and enrich the resource, as well as to optimize it through the revision of inflectional paradigms currently implemented. The objective of the present work is therefore to increase the number of correctly recognized verbs from 37 to approximately 800. To support this research, the corpus of Polish texts has also been expanded through the inclusion of public-domain works, reaching a total size of nearly two million words.
The complexity of a verbal inflectional paradigm depends largely on how the notion of paradigm is defined, which in any case is tied to the theoretical framework we adopt (see e.g. [3, 5]). In general terms, a linguistic paradigm may be understood as a pattern or template used to generate word forms from a given lemma. Determining which forms belong to a paradigm—verbal, for instance—requires engaging with the distinction between inflection and derivation, in particular the question of which forms should be considered part of verbal inflection and which should be treated as independent words derived from verbs [6]. We believe that the NooJ

formal grammar, like any formal grammar, may influence our decisions in this area, with efficiency and economy of the grammar in mind.

The inflectional model proposed by Bogacki (see NLR) includes the following sets of grammatical word forms: tenses (past, present and futur), moods (e.g. indicative, imperative and conditional), impersonal forms (the so-called Polish *bezosobnik*), adverbial participles (perfect and imperfect), and adjectival participles (active and passive). A set, that could potentially be included in the model, is the gerund (verbal noun), since it inherits the verb's aspect (*robić* [to do] > *robienie* [act of doing], *zrobić* [to have done] > *zrobienie* [act of having done]), as well as its argument structure (e.g. *robić*(x,y) > *robienie*(x,y)). Following Saloni's conjugation tables [8], the model could also include the so-called perfect adjectival participle (e.g. osiawiały [gone gray]), a category less frequently recognized as part of the verbal paradigm [6].

With around 300 forms that can be generated to complete a single paradigm, we may thus be dealing with a highly complex system of verbal inflection. However, not all Polish verbs admit all the sets of forms mentioned above. The acceptability of a given form depends, first, on the verb's aspect; for example, the imperfective verb robić allows present tense forms, whereas the perfective verb *zrobić* do not. Consequently, robić allows imperfect adverbial participle (*robiąc* [while doing]), whereas zrobić allows perfect adverbial participle (*zrobiwszy* [having done]). Neither accepts the inverse participle accepted by the other (**zrobiąc*, **robiwszy* are not allowed).

Other factors, such as (in)transitivity and reflexivity, may also determine the acceptability of certain forms. These considerations motivate the optimization of the existing model in order to maximize verb coverage while avoiding redundant patterns with highly similar morphological characteristics that differ, however, in properties such as aspect, (in)transitivity, and reflexivity, which in turn prevent the generation of certain semantically incompatible word forms.

Keywords: NooJ, verbs of motion, inflection versus derivation, aspect, Polish verb.

References

[1] Krzysztof Bogacki and Ewa Gwiazdecka. 2012. Derivational Structure of Polish Verbs and the Expansion of the Dictionary. In K. Vučković, B. Bekavac & M. Silberztein (Eds.), Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference. Cambridge: Cambridge Scholars Publishing.

[2] Krzysztof Bogacki and Ewa Gwiazdecka. 2014. Słowniki polskie dla platformy NooJ. Prace Filologiczne, vol. LXV, pages 11–28.
[3] Geert Booij. 1993. Against split morphology. In Yearbook of Morphology, 23. Dordrecht / Boston / London: Kluwer Academic Publishers, pages 27–49.
[4] Agnieszka Dziob, Maciej Piasecki, and Ewa Rudnicka. 2019. plWordNet 4.1 – a Linguistically Motivated, Corpus-based Bilingual Resource. In Proceedings of the 10th Global Wordnet Conference, pages 353–362.
[5] Adam Heinz. 1961. Fleksja a derywacja. Język Polski, 41, pages 343–354.
[6] Krystyna Kowalik. 2024. Fleksja. Przewodnik językowo-encyklopedyczny po gramatyce semantycznej języka polskiego w ujęciu historycznym. Instytut Języka Polskiego PAN. URL: https://gramsem.ijppan.pl/hasla/fleksja/
[7] NooJ Linguistic Resources (Polish), https://nooj.univ-fcomte.fr/resources.html
[8] Zygmunt Saloni. 2010. Czasownik polski. Warszawa: Wiedza Powszechna.
[9] Max Silberztein. 2003. NooJ Manual. http://www.nooj4nlp.net

From coded lexicon to NooJ dictionary: property-based annotation of ecological lexical innovation

Lara Nougaret
Université de Montpellier Paul-Valéry, Montpellier, France

This paper reports on a doctoral research program in language sciences (terminology, lexicology and discourse analysis) addressing lexical innovation in contemporary ecological discourse. Earlier outputs from this program included the design of NooJ definitional grammars (presented at NooJ Week 2025; [3]). The present paper complements that line of work by implementing a manually coded terminological inventory as an annotated NooJ dictionary for French, enabling corpus annotation and property-based querying. In line with research on neology and environmental discourse emphasizing the productivity and rapid circulation of new forms ([2]; [1]), we operationalize ecological lexical innovation through a reusable NooJ lexical resource.

The project builds a large inventory of emerging nominations and multi-word units, compiled into a lexical database and manually categorized using a terminological-discursive coding scheme. Each dictionary entry (single-word item or multiword unit) is enriched with lexical properties capturing: (i) semantic reference, (ii) entity type, (iii) discursive orientation, (iv) origin domain, and (v) circulation field, with metadata such as axiological load, productive morphological markers (prefixes/suffixes), or rhetorical figures. This encoding shifts corpus exploration from surface-string searches to category-driven observation, allowing targeted concordance extraction and quantitative summaries by property.

A key technical choice concerns multiword ecological terminology: multiword units are integrated as dictionary entries and systematically marked with +UN-AMB, ensuring that NooJ prioritizes the lexicalized unit over compositional analyses. This is crucial in a domain where innovation frequently takes the form of productive noun phrases and series. The contribution is both methodological and resource-oriented: it provides a reproducible workflow for converting a manually coded terminological inventory into a NooJ lexical resource and lays the groundwork for further NooJ developments based on the same property layer.

Keywords: Lexicology, Discourse Analysis, Ecology.

References

[1] V. Balnat and C. G´erard, editors. Neologica 2022, n° 16: N´eologie et environnement. Neologica, 2022.

[2] A. Jackiewicz and O. Kraif. R´ealit´es ´emergentes, possibles, d´esirables: Comment les nommons-nous? Humanités environnementales: sciences, arts et citoyenneté face aux changements globaux, 1(1), 2021.

[3] A. Jackiewicz and L. Nougaret. Approcher la notion de commun(s) par l'exploration de corpus ´ecologique(s). In Semaine NOOJ 2025, 2025.

[4] M.-C. L'Homme. Sur la notion de terme . Meta, 50(4):1112–1132, 2005.

[5] J.-F. Sablayrolles. Nomination, d´enomination et n´eologie: Intersection et diff´erences sym´etriques. Neologica: revue internationale de la n´eologie, (1):87–99, 2007.

[6] J.-F. Sablayrolles. Comprendre la ńeologie: Conceptions, analyses, emplois. Lambert-Lucas, 2019.

Interactive Web Interface for Learning Ukrainian with NooJ

Victor Rabiet[1], **Divna Petković**[2]
[1] École Normale Supérieure, Paris, France
[2] NYU, Paris, France

This contribution presents the linguistic design, pedagogical objectives, and technical implementation of an interactive web interface dedicated to Ukrainian, developed on top of NooJ resources. The tool is intended both as a didactic environment for learners of Ukrainian and as a platform for creating and exploring linguistically annotated sentence corpora.

The interface is web-based and designed for a high level of interactivity. Users can access linguistic information directly from the text through simple actions such as hovering over a word or selecting options via check boxes. The system provides structural information as well as detailed grammatical descriptions for parts of speech. For inflected words, full declension or conjugation paradigms are displayed in a separate panel adjacent to the text, allowing users to move seamlessly between contextualized usage and paradigmatic representation.

Particular attention is given to verbal aspect, a notion that is notoriously difficult for non-Slavic speakers. For each verb, the interface encodes the aspectual pair and provides illustrative xamples. At a basic level, Ukrainian aspect is often introduced through the opposition between perfective verbs, which express the completion of an action within a given timeframe, and imperfective verbs, which express duration without specifying completion. While this distinction is valid in this context, it remains only a partial account of aspectual meaning in Ukrainian, which encompasses a wide range of semantic nuances (cf. Chinkarouk [1]). Building on previous work by Saint-Joanis [2][3] and Silberztein [3][4], which provides both a formal framework for modeling aspectual relations and a large-scale electronic dictionary in NooJ (over 13 000 verb entries and more than 300 morphological paradigms), the proposed interface aims to make these nuances observable in a systematic and pedagogically accessible way.

By combining fine-grained linguistic formalization with an interactive and learner-oriented interface, this project seeks to bridge the gap between computational linguistic resources and concrete classroom needs, offering instructors and students a practical tool for exploring Ukrainian grammar in context.

References

[1] Chinkarouk, Oleg. *Système verbal et aspect en ukrainien, en russe et en français: étudecontrastive.* Diss. Université Stendhal (Grenoble; 1970-2015), 1996.

[2] Saint-Joanis, Olena. *Formalisation de la langue ukrainienne avec NooJ: préparation du module ukrainien*. Diss. Université Bourgogne Franche-Comté, 2024. https:// theses.hal.science/tel-04690884/

[3] Max Silberztein, Olena Saint-Joanis. *Formalisation de la relation entre les verbes imperfectifs et perfectifs en ukrainien*. Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, Jun 2022, Lille, France.

[4] Silberztein, Max*. La formalisation des langues: l'approche de NooJ*. ISTE Group, 2015.

The Discourse of Dubbing: Balancing Art and Technique in Professional Identity – A NooJ-Aided Analysis

Maïa Joubert
Université de Montpellier Paul-Valéry, Montpellier, France

This paper is part of a PhD research project in Linguistics (Discourse Analysis), conducted at the ReSO laboratory under a CIFRE fellowship in partnership with CTM Solutions / TITRAFILM, a French dubbing company. It proposes a NooJ-aided linguistic analysis of a corpus of interviews dedicated to the professional roles within the audiovisual dubbing industry.
The corpus consists of 20 semi-structured interviews conducted with five professional categories involved in the dubbing production chain (four participants per category): voice actors, artistic directors, dubbing scriptwriters (adapters), spotters (detectors), and sound engineers. All interviews have been fully transcribed, representing approximately 350 pages of textual data. The interviews were structured around four thematic axes: career paths and entry into the industry; work practices; interprofessional collaborations; and perceptions of sectoral transformations. The study adopts a qualitative discourse analysis perspective, focusing on the way professionals discursively construct their identity and describe their practices.
Initial close readings of the corpus, particularly within the discourse of actors and artistic directors, reveal a recurring lexical field of emotion linked to artistic practices. Simultaneously, terms referring to technical constraints occupy an equally central position. These observations suggest that the interplay between the artistic and technical dimensions constitutes a structuring tension within the professional world of dubbing.

Formalizing Professional Discourse through NooJ Lexical Grammars
The methodological contribution of this paper concerns the use of NooJ to formalize and explore this tension at the lexical level. Two lexical grammars will be developed: one focusing on terms related to the emotional dimension, and the other on terms linked to technical constraints (synchronization, timing, recording). These grammars will then be applied to the entire corpus to observe their distribution across professional categories and to analyze how artistic and technical lexicons coexist and intersect within the discourse. For instance, we will test the hypothesis that the lexical field of emotion effectively maps onto the artistic dimension of professional practices.

Particular attention will also be paid to metalinguistic reformulations. A lexical grammar of dubbing-specific jargon (e.g., "*bande rythmo*," "*rec*," "*croisillé*," "*détecteur*," "*adaptateur*") will be constructed to identify segments where professionals explain or redefine these specialized terms. The aim is to examine how industry stakeholders discursively construct and negotiate their professional knowledge, and whether these definitions converge or vary according to the specific trade. In other words, the objective is to linguistically model the discursive sites where the balance between the technical and the aesthetic—the core of professional identity in the dubbing sector—is negotiated.

This communication aims to demonstrate how NooJ can be mobilized to build and apply lexical resources tailored to a specialized professional corpus, and how linguistic tools support an analysis situated at the interface of linguistics, professional discourse analysis, and digital humanities within a collaborative research context.

The Representation of the Black Lives Matter Movement in US Media: A Comparative Lexical Analysis

Francesco Ugo Capitelli, Rosa Santarsiere and **Alessia Santonicola**
Università di Napoli L'Orientale, Naples, Italy

The Black Lives Matter (BLM) movement has represented a turning point in contemporary social history, triggering a global debate on systemic inequality. However, journalistic coverage of the movement in the United States has been characterized by deep political polarization1,2. The main objective of this project is to investigate this polarization through a lexical analysis of negative adjectives. Specifically, the study aims to quantify and compare how terms related to violence, disorder, criminality, and aggression are deployed in conservative and progressive media to frame the protests and attribute responsibility. The analysis is based on a balanced corpus of 60 online news articles, collected between May 2020 and January 2021. The dataset consists of 30 articles from six conservative outlets (e.g., New York Post, The Washington Times) and 30 articles from six progressive/liberal outlets (e.g., CNN, Democracy Now). To perform the analysis, a lexical resource was developed within the NooJ linguistic development environment: a dictionary of negative adjectives was created by filtering a high-frequency word list and applying a custom inflectional grammar to capture morphological variations (e.g., comparative and superlative forms). NooJ was utilized to process the corpus, extract concordances, and calculate advanced statistics, specifically the Z-Score, to measure the deviation of negative adjective density from the corpus mean.

This computational phase was complemented by a manual validation of the concordances to identify the semantic referents of the adjectives (e.g., Protesters/BLM vs. Police/Authorities), allowing for a precise mapping of how blame is targeted. The results reveal an imbalance in the media representation of the movement. The conservative corpus demonstrates a systematic over-representation of negative adjectives (with Z-scores frequently exceeding +2), where the majority (65.1%) targets the BLM protesters. On the other hand, the progressive corpus shows a statistical under representation of lexical negativity (negative Z-scores); significantly, when negative adjectives are used, they are often redirected toward law enforcement and the institutional response (28.9%), shifting the blame from the protesters to the system. The analysis highlights that the adjective "violent" is the most polarizing lexical unit, appearing with a frequency four times higher in conservative texts. These findings demonstrate that political

orientation does not only influence opinion but builds different realities. Future extensions of this work may include expanding the dataset to Big Data in order to perform an analysis on a larger scale and integrate other morphological categories (verbs and nouns). Furthermore, the resource developed could provide tools to help users decode hidden biases and hate speech in journalistic narratives.

Keywords: Black Lives Matter, Media Polarization, Lexical Analysis, NooJ, Negative Adjectives, Protest Paradigm.

References

[1] Cox, J. B. 2022. Black Lives Matter to media (finally): A content analysis of news coverage during summer 2020. Newspaper Research Journal, 43(2), pages 155-175. https://doi.org/10.1177/07395329221092719

[2] Kim, Sei-Hill, Zdenek Rusek Kotva, Ali Zain, and Yu Chen. 2024. Black Lives Matter and Partisan Media. Journalism and Media, 5, pages 78-91. https://doi.org/10.3390/journalmedia5010006

Prodotto presso il Centro Pleiadi
Sezione Officine Grafico-Editoriali – Il Torcoliere