

Preface

NooJ is a linguistic development environment that provides tools for linguists to construct linguistic resources that formalize a large gamut of linguistic phenomena: typography, orthography, lexicons for simple words, multiword units and discontinuous expressions, inflectional and derivational morphology, local, structural and transformational syntax, and semantics.

For each resource that linguists create, NooJ provides parsers that can apply it to any corpus of texts and extract examples or counter-examples, annotate matching sequences, perform statistical analyses, and so on. NooJ also contains generators that can produce the texts that these linguistic resources describe, as well as serving as a rich toolbox that allows linguists to construct, maintain, test, debug, accumulate, and reuse linguistic resources.

For each elementary linguistic phenomenon to be described, NooJ proposes a set of computational formalisms, the power of which ranges from very efficient finite-state automata to very powerful Turing machines. This makes NooJ's approach different from most other computational linguistic tools that typically offer a unique formalism to their users.

Since its first release in 2002, NooJ has been enhanced with new features every year. Linguists, researchers in social sciences, and more generally all professionals who analyze texts have contributed to its development and participated in the annual NooJ conference. In 2013, a new version of NooJ was released, based on the JAVA technology and available to all as an open source GPL project. Moreover, several private companies are now using NooJ to construct business applications in several domains, from business intelligence to opinion analysis.

At the opening of the International NooJ 2016 Conference, which was held during June 9–11 at the University of South Bohemia in České Budějovice (Czech Republic), Max Silberstein presented his book *Formalizing Natural Languages: The NooJ Approach* (Wiley, 2016), which describes the theoretical and methodological backgrounds of the NooJ approach. The present volume contains 21 articles selected from the other 32 papers presented at the conference. These articles are organized in four parts: “Vocabulary and Morphology” containing seven articles; “Corpus Processing and Information Extraction” containing six articles; “Syntactic Analysis” containing five articles; and “Semantic Analysis and Its Applications” containing three articles.

The articles in the first part cover the construction of electronic dictionaries, the description of phonetic and morphological information, and the construction of bilingual electronic dictionaries that can be used by machine translation software.

- Hamid Annouz's article “Treatments of the Kabyle Derived Nominal Verbs with NooJ” describes the formalization of nominalization of Kabyle verbs in the context of a full-coverage description of the Kabyle vocabulary.

- Stanislaw Lysy, Hanna Stanislavenka, and Yury Hetsevich’s article “Addition of IPA Transcription to the Belarusian NooJ Module” shows how to integrate phonetic information (including stress position) in a NooJ dictionary.
- Kristina Kocijan, Marijana Janjić, and Sara Librenjak’s article “Recognizing Diminutive and Augmentative Croatian Nouns” presents a set of morphological grammars used to recognize Croatian diminutive and augmentative nouns, using the base nouns listed in the Croatian dictionary.
- Dhekra Najar, Slim Mesfar, and Henda Ben Ghezela’s article “Inflectional and Morphological Variation of Arabic Multi-Word Expressions” presents a linguistic module that includes dictionaries, morphological and syntactic grammars, used to recognize multi-word units in Arabic texts.
- Maximilian Duran’s article “Quechua Module for NooJ Multilingual Linguistic Resources for MT” presents the first available electronic dictionary for Quechua that formalizes inflectional and derivational morphology.
- Francesca Esposito and Annibale Elia’s article “NooJ Local Grammars for Innovative Start-Up Language” presents a system capable of analyzing the vocabulary of specialized texts such as the documents of start-up companies.
- Hager Cheikhrouhou’s article “Arabic Translation of the French Auxiliary: Using the Platform NooJ” uses the class “X” of auxiliary verbs from the Dubois LVF dictionary and presents the corresponding dictionary for Arabic verbs.

The articles in the second part involve the construction of automatic software capable of parsing and extracting meaningful information from texts by retrieving terms or named entities automatically:

- Nadia Ghezaiel Hammouda and Kais Haddar’s article “Integration of a Segmentation Tool for Arabic Corpora in NooJ Platform to Build an Automatic Annotation Tool” presents a set of grammars used to recognize and annotate sentences from Arabic texts.
- Yury Hetsevich, V. Varanovich, E. Kachan, I. Reentovich, and S. Lysy’s article “Semi-automatic Part-of-Speech Annotating for Belarusian Dictionaries Enrichment in NooJ” presents an algorithm that uses a corpus of one million words to automatically enrich the NooJ Belarusian dictionary.
- Francesca Parisi’s article “Clinical Term Recognition: From Local to LOINC Terminology. An Application for Italian Language” uses a clinical and biological terminological database to associate texts with a specific code from the LOINC standard.
- Walter Koza’s “Enumerative Series in Spanish: Formalization and Automatic Detection” presents a set of syntactic local grammars used to recognize and process enumerations in Spanish.
- Mohamed Aly Fall Seideh, Hela Fehri, and Kais Haddar’s article “Recognition and Extraction of Latin Names of Plants for Matching Common Plant Named Entities” uses the International Code of Botanical Nomenclature as an interface to build an automatic French–Arabic term translation system.
- Hiba Chenny and Slim Mesfar’s article “Generating Alerts from Automatically Extracted Tweets in Standard Arabic” presents a text parser capable of mining texts to extract relevant information from tweets written in Arabic.

The articles in the third part describe the construction of sophisticated syntactic grammars and the use of such grammars by automatic paraphrasing generators built with NooJ:

- Krešimir Šojat, Božo Bekavac, and Kristina Kocijan’s article “Detection of Verb Frames with NooJ” presents a parser capable of automatically recognizing in Croatian texts derived verbs using morphological grammars, and then associate each recognized form with its syntactic valency, thanks to a chunker that uses a set of syntactic local grammars.
- Peter A. Machonis’s article “Phrasal Verb Disambiguation Grammars: Cutting Out Noise Automatically” presents a set of English dictionaries and grammars used to automatically disambiguate sequences of texts that may (or may not) represent phrasal verb constructions.
- Mario Monteleone’s article “NooJ Local Grammars for Endophora Resolution” presents the different types of endophora, and proposes several techniques using NooJ local grammars to solve them.
- Alberto Maria Langella’s article “Paraphrases for the Italian Communication Predicates” presents an Italian transformational grammar capable of producing paraphrases of sentences that contain a communication predicate.
- Cristina Mota, Anabela Barreiro, Francisco Raposo, Ricardo Ribeiro, Sérgio Curto, and Luísa Coheur’s article “eSPERTo’s Paraphrastic Knowledge Applied to Question-Answering and Summarization” presents the eSPERTo paraphrastic engine and its use to produce summaries of Portuguese texts.

The articles in the last part of this volume describe business applications built with NooJ that use deep semantic analysis, using complex syntactic and semantic dictionaries and grammars.

- Maria Pia di Buono’s article “Endpoint for Semantic Knowledge (ESK)” presents the ESK framework and shows how NooJ can be used to process semantic information using its dictionaries and grammars.
- Francesca Esposito and Maddalena della Volpe’s article “Using Text Mining and Natural Language Processing to Support Business Decision: Towards a NooJ Application” presents an automatic decision-making system capable of parsing business documents using specialized linguistic resources.
- H  la Fehri, Mohamed Aly Fall Seideh, and Sondes Dardour’s article “A Decision-Support Tool of Medicinal Plants Using NooJ Platform” presents an automatic system capable of recommending medicinal plants based on the recognition of various criteria present in a French text such as the age or symptoms.

This volume should be of interest to all users of the NooJ software because it presents the latest development of the software as well as its latest linguistic resources. To date, there are NooJ modules available for over 50 languages; more than 3,000 copies of NooJ are being downloaded each year.

Linguists as well as computational linguists who work on Arabic, Belarusian, Chinese, Croatian, English, French, Italian, Kabyle, Portuguese, Spanish, or Quechua will find advanced, up-to-the-minute linguistic studies for these languages in this volume.

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

February 2017

Linda Barone
Max Silberztein
Mario Monteleone

Automatic Processing of Natural-Language Electronic
Texts with NooJ

10th International Conference, NooJ 2016, České
Budějovice, Czech Republic, June 9-11, 2016, Revised
Selected Papers

Barone, L.; Monteleone, M.; Silberztein, M. (Eds.)

2016, XII, 259 p. 155 illus., Softcover

ISBN: 978-3-319-55001-5