WORKSHOP

NOOJ ENVIRONMENT FOR RULE-BASED NATURAL LANGUAGE PROCESSING

INSTRUCTOR: MARIO MONTELEONE (https://docenti.unisa.it/005793/home)

NooJ is a linguistic development environment software as well as a corpus processor constructed by **Max Silberztein (ELLIAD, Université de Franche-Comté, France, http://elliadd.univ-fcomte.fr/fiches/silberzteinmax)**. It firstly originated in investigations by Silberztein and the INTEX community of linguist users into the Lexicon-Grammar approach of Maurice Gross' (LADL, Université Paris 7, France, https://www.wikiwand.com/en/Maurice_Gross), which states that no grammar rule can be developed independently from a strict delimitation of its domain of application.

NooJ has been used as a corpus processor by researchers in Linguistics, History, Psychology, Literature studies, Sentiment Analysis projects, Data Mining, and for processing musical notes and sentences.

NooJ allows its users to:

1. Construct the four classes of the Chomsky-Schützenberger hierarchy of generative grammars: Finite-State Grammars, Context-Free Grammars, Context-Sensitive Grammars as well as Unrestricted Grammars, using either a text editor (e.g. to write down regular expressions), or a Graph editor;
2. Develop orthographical and morphological grammars, finite-state transducer dictionaries of simple words, of compound words as well as discontinuous expressions, local syntactic grammars (such as Named Entities Recognizers), structural syntactic grammars (that produce syntactic trees) as well as Zellig Harris' transformational grammars.

All NooJ parsers process Atomic Linguistic Units (ALUs), as opposed to word forms (i.e. sequences of letters between two space characters). This allows NooJ syntactic parser to parse sequences of word forms such as “can not” in the same way as contracted word forms such as “cannot” or “can't”.

ALUs are represented by annotations that are stored in a Text Annotation Structure (or TAS): all NooJ parsers add, or remove annotations in the TAS. A typical NooJ analysis involves applying to a text a series of elementary grammars in cascade, in a bottom-up approach (from spelling to semantics).

NooJ applies grammars to texts in linear time, as most NooJ Context-Free Grammars can often be derecursived. NooJ Context-Sensitive Grammars are made of two parts: the first is a Context-Free (or even a Finite-State Grammar) applied to texts, the second consists in a set of constraints applied to matching sequences, each one performed in constant time. NooJ unrestricted grammars are context-sensitive grammars that can contain variables and can modify the text input. They are typically used to perform transformational analysis and generation. Furthermore, when used in conjunction with multilingual lexicons, they can be used to perform Machine Translation.

WORKSHOP ACTIVITIES AND LEARNING OBJECTIVES

Activities:

- Introduction to NooJ
- Text and Corpora Analysis
- Dictionaries and Vocabularies Description
- Grammar Building
- NooJ Local Grammars
- Structural and Transformational Analysis

After the theoretical illustrations, for each activity practical exercises will be held.

Learning Objectives:
Introduce Workshop participants to:

- Rule-based NLP procedures and techniques;
- NooJ as a language-development environment for text processing and querying, concordance building and analysis;
- NooJ six types of linguistic resources (typography and spelling, lexical morphology, dictionaries, inflectional morphology, derivative morphology, productive morphology);
- NooJ (Local) Grammar construction and application;
- NooJ structural syntax (syntax tree and annotations) and transformations (paraphrases, semantic analysis, translation).

REQUIREMENTS

- Basic knowledge of General Linguistics (mainly morphology and syntax);
- NooJ environment installed on a Windows notebook (NooJ download page: https://www.nooj-association.org/downloads.html) plus the download of the linguistic resources of one's mother tongue (download page: https://www.nooj-association.org/resources.html);
- Download and reading of the NooJ Manual (http://www.nooj-association.org/files/NooJManual.pdf).

EXTRA MATERIALS

No extra material will be necessary for participation in the Workshop. Furthermore, all analyzes and exercises carried out will remain stored on the computers of the individual participants.

MAIN REFERENCES

Max Silberztein, *Formalizing Natural Languages: the NooJ Approach*. 2016. Wiley Eds. Hoboken, NJ, USA

• For more articles, case studies and scientific materials: https://www.nooj-association.org/references.html.

WORKSHOP WEBSITES

https://www.nooj-association.org/index.html