

NooJ Workshop: A Sophisticated Finite-State Linguistic Analysis Tool for Corpora

Max Silberztein and Tamás Váradi¹

Abstract

The aim of this 2-hour workshop is to give an overview of the NooJ corpus processing tool. The workshop will focus on the linguistic aspects of corpus annotation and querying.

NooJ is a comprehensive linguistic development environment written for the Windows .NET platform. On the one hand, it allows linguists to construct large-coverage linguistic resources in the form of dictionaries and grammars, and provides tools for their maintenance: contracts, debuggers, etc. On the other hand, it is a corpus processor that can apply these resources to large texts in order to annotate them, build sophisticated concordances, analyse various syntactic and semantic phenomena, and retrieve and extract information from them.

The main attraction of NooJ for the ordinary corpus linguist is the ease with which the complex functionalities can be handled. This is particularly evident in the graphical interface through which sophisticated cascading finite-state grammars can be developed. The finite-state capabilities are substantially enhanced through the use of variables, typed features and inheritance mechanisms etc.

Format and content

The workshop will consist of four 25-minute presentations and an interactive "how-to" session.

– The first presentation (by Max Silberztein) will be an overview of the design philosophy and the architecture of the system, and then a run-down of how NooJ can be applied in the daily routine tasks of corpus analysis: how to import a corpus and annotate it, how to build complex concordances from simple and complex queries, and how to export results in an XML document;

– The second presentation (by Tamas Varadi) will focus on the linguistic capabilities of NooJ's local grammars that are enhanced finite-state transducers using features and variables. Issues discussed will also include the writing, compiling and prioritized deployment of dictionaries and grammars, and disambiguation at various levels.

– The third presentation (by Kata Gabor) will present a robust rule-based syntactic parser for Hungarian built with NooJ. Hungarian produces a challenge for parsers because of its extremely complex morphology (with the inevitable heavy morphosyntactic ambiguity) coupled with free constituent order within the clause. The syntactic parser is composed of a set of cascaded local grammars which makes extensive use of NooJ's special features such as variables, feature percolation, lexical

¹ Université Franche-Compte, Hungarian Academy of Sciences
e-mail: tavaradi@gmail.com

constraints and structured lexical resources for capturing even long-distance dependency relations. The presentation will outline the architecture of the grammar as well as its application in annotating corpora.

– The last presentation (by Slim Mesfar) will demo how to quickly build an information extractor: The last presentation (by Slim Mesfar) will demo how to quickly build an information extractor:

The session will conclude with an interactive part during which members of the audience will be encouraged to raise "how-to do it?" type of questions, which the presenters will try to answer on the spot, through quick demonstration, if possible.

Justification

While NooJ has been used in an impressive number of languages ranging from French to Hungarian and Arabic, it is relatively little known among the English speaking corpus linguistic community. The workshop will be a good opportunity to introduce the benefits of this freely available tool to the corpus linguist.

Audience participation

The presentations will make every attempt to strike a balance between being sufficiently informative and at the same time remaining accessible to everyone without any previous familiarity with the system. The 25 minute presentation slots will serve to raise interest for the closing interactive part of the workshop where members of the audience will be encouraged to raise "How to" type of questions to give them a feel of how NooJ might be employed for their own needs.