Corpus Analysis with NooJ

Monday, May 20, 2012

Presenters:

Kristina Vučkovic, Tamás Váradi, Max Silberztein

Tutorial Programme/Overview

NooJ is a freeware language-engineering development environment used to formalize and integrate nine levels of linguistic phenomena: orthography and typography, lexical, inflectional and derivational morphology, local, structural and transformational syntax, semantics.

For each of these levels, NooJ provides linguists with one or more formal framework specifically designed to facilitate the description of each phenomenon, as well as parsing, development and debugging tools designed to be as computationally efficient as possible, from Finite-State to Turing machines. This approach distinguishes NooJ from other computational linguistic frameworks that provide a unique formalism that is supposed to cover all linguistic phenomena.

As an Engineering development environment, NooJ contains tools to help construct, test, debug, maintain and accumulate large sets of linguistic resources, as well as tools to process large texts and corpora. The system has been developed since 2002 and it has been used to build over 20 language modules.

As a corpus processing tool, NooJ allows researchers in various social sciences to extract information from any text or corpus (i.e. not tagged) by applying sophisticated queries based on concepts rather than word forms and build indices and concordances, automatically annotating texts, perform statistical analyses on concepts, etc.

NooJ is freely available, it runs on Windows, Linux, Solaris and Mac OSX Lion. Linguistic modules can already be freely downloaded for over a dozen languages. See www.nooj4nlp.net for more information on NooJ; the page "**doc & help**" provides references to NooJ-related publications.

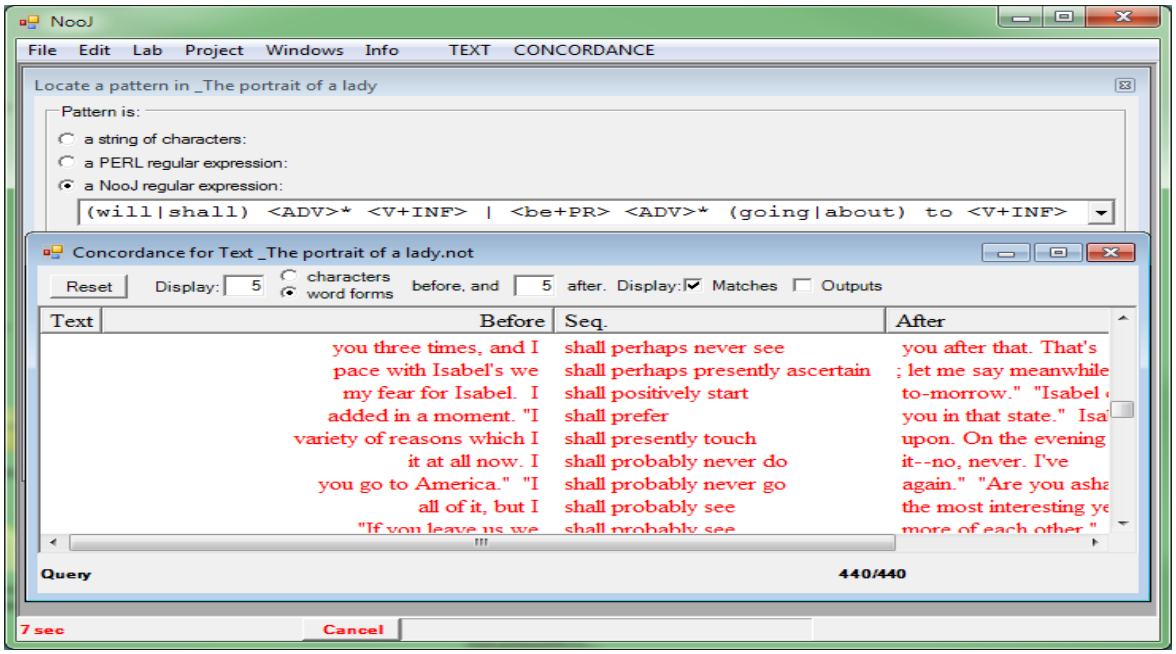
This tutorial intends to help participants to master three basic NooJ functionalities: corpus processing, formalization of linguistic units, syntactic parsing and the automatic annotation of texts.

Tutorial Description/Outline/Contents

The tutorial session consists of three labs and one presentation. Participants must come with their laptop with NooJ v3 (either .NET or MONO) pre-installed.

1. Corpus Processing (Kristina Vučkovic, Univ. of Zagreb)

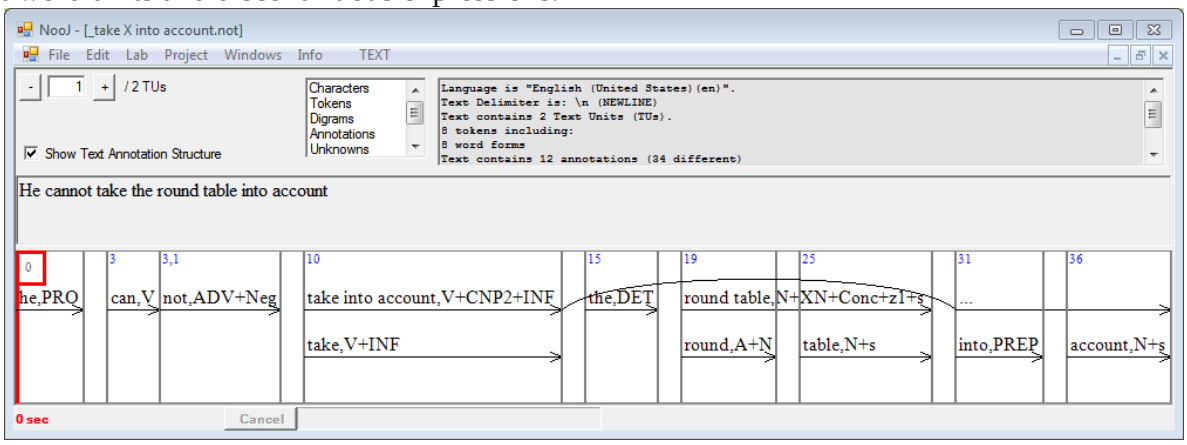
Import a text in any file format (including XML) and build a corpus. A corpus is a collection of text files that will share the same linguistic resources. Apply simple and complex queries to texts in order to build concordances. Perform statistical analyses on any query.



A concordance

2. **Linguistic Units, Lexicons and morphology** (Max Silberztein, Univ. of Franche-Comté)

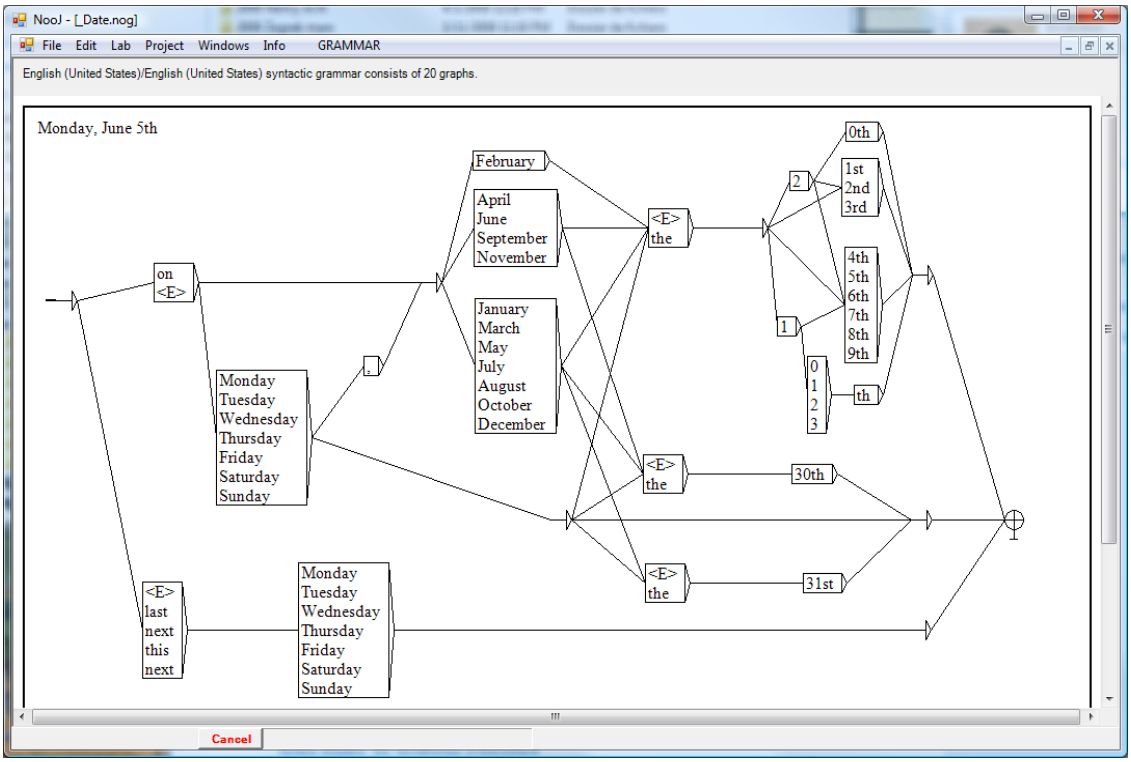
NooJ's basic objects are Atomic Linguistic Units (ALU). We will learn how to construct dictionaries and associate them with morphological paradigms. When parsing a text, NooJ represents all ALUs in the Text Annotation Structure: affixes (inside word forms), simple words, multiword units and discontinuous expressions.



The Text Annotation Structure

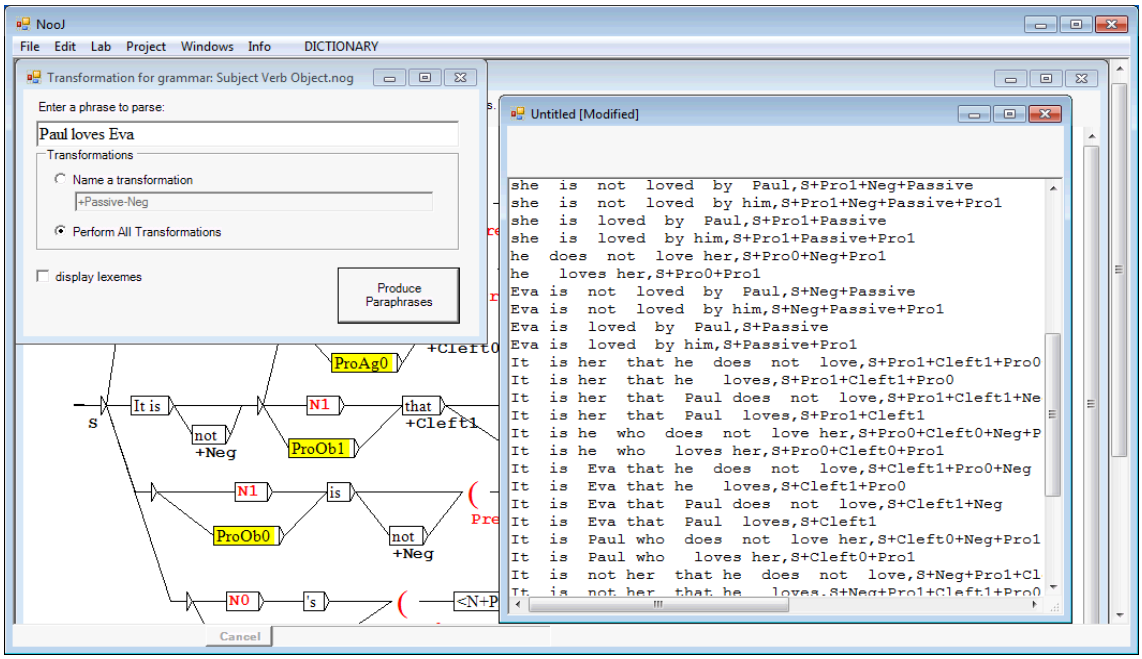
3. **Local Grammars** (Max Silberztein, Univ. de Franche-Comté)

What is a local grammar? Develop and combine local grammars. Apply local grammars to texts.



A local grammar

Maintain and merge grammars; debug grammars with the debugger.



Automatic paraphrasing of a sentence

4. Conclusion : Best Practices from the CESAR Project (Tamás Váradi, Budapest Academy of Sciences)

The METANET CESAR Project: constructing large-coverage description of linguistic phenomena for Eastern European Languages. Building an Open Source JAVA toolbox based on NooJ technology.