LREC2012 NooJ Tutorial

Outline

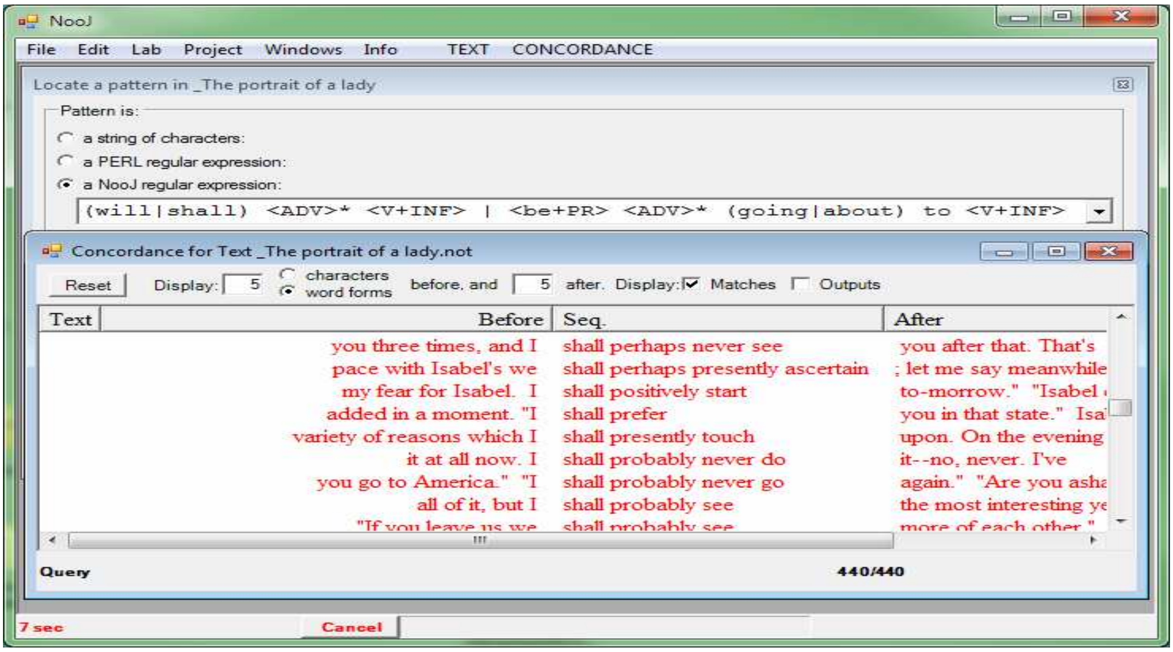
Morning session – Monday, 21 May 2012

NooJ is a freeware language-engineering development environment for formalizing and integrating nine levels of linguistic phenomena: orthography and typography, lexical, inflectional and derivational morphology, local, structural and transformational syntax, semantics. 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, including language modules developed in the framework of the META-NET CESAR ICT-PSP project. This tutorial intends to help participants to master four basic NooJ functionalities: corpus processing, formalization of linguistic units, syntactic parsing.

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

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

Import a text in any file format. Need to define NooJ text units, which are independent areas inside which linguistic resources are applied. 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.



Locate a pattern in _The portrait of a lady

Pattern is:

○ a string of characters:
○ a PERL regular expression:
● a NooJ regular expression:

```
(will|shall) <ADV>* <V+INF> | <be+PR> <ADV>* (going|about) to <V+INF>
```

Concordance for Text _The portrait of a lady.not

Display: 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches □ Outputs

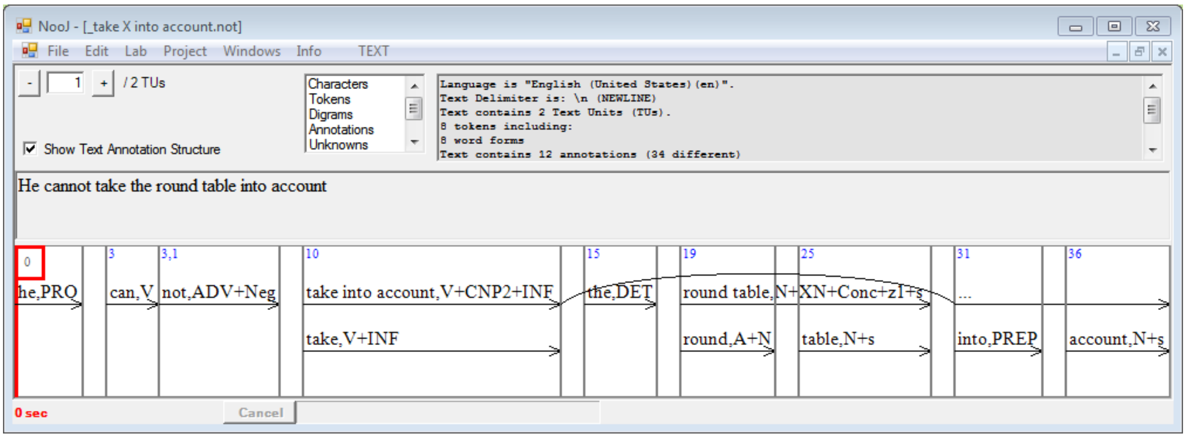
Text	Before	Seq.	After
	you three times, and I	shall perhaps never see	you after that. That's
	pace with Isabel's we	shall perhaps presently ascertain	; let me say meanwhile
	my fear for Isabel. I	shall positively start	to-morrow." "Isabel
	added in a moment. "I	shall prefer	you in that state." Isa
	variety of reasons which I	shall presently touch	upon. On the evening
	it at all now. I	shall probably never do	it--no, never. I've
	you go to America." "I	shall probably never go	again." "Are you asha
	all of it, but I	shall probably see	the most interesting ye
	"If you leave us we	shall probably see	more of each other."

Query 440/440

7 sec

2. Linguistic Units and annotations (Max Silberztein, Univ. of Besançon)

NooJ's basic objects are Atomic Linguistic Units (ALU). NooJ contains tools to formalize nine levels of linguistic phenomena. NooJ manages annotations (rather than tags) to make its different parsers communicate and keep unresolved ambiguities in a Text Annotation Structure.


NooJ - [_take X into account.not]
File Edit Lab Project Windows Info TEXT
- 1 + / 2 TUs
☑ Show Text Annotation Structure
Characters
Tokens
Digrams
Annotations
Unknowns
0 sec
Cancel

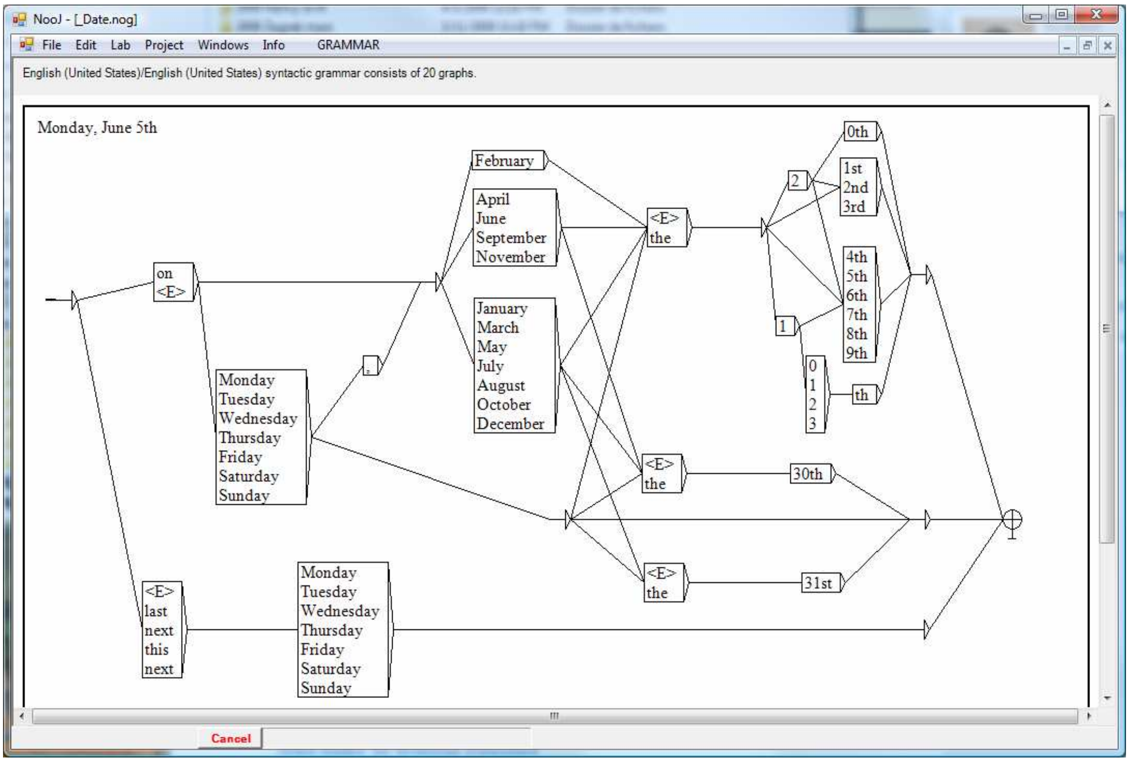
```
Language is "English (United States)(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
8 tokens including:
8 word forms
Text contains 12 annotations (34 different)
```

He cannot take the round table into account

3. Local Syntax (Tamas Varadi, Budapest Academy of Sciences)

What is a local grammar? Developing and combining local grammars. Applying local grammars in cascade. How to maintain grammars with the help of NooJ debugger and contracts.

Automatic annotation of linguistic units and named entities. XML import and export.

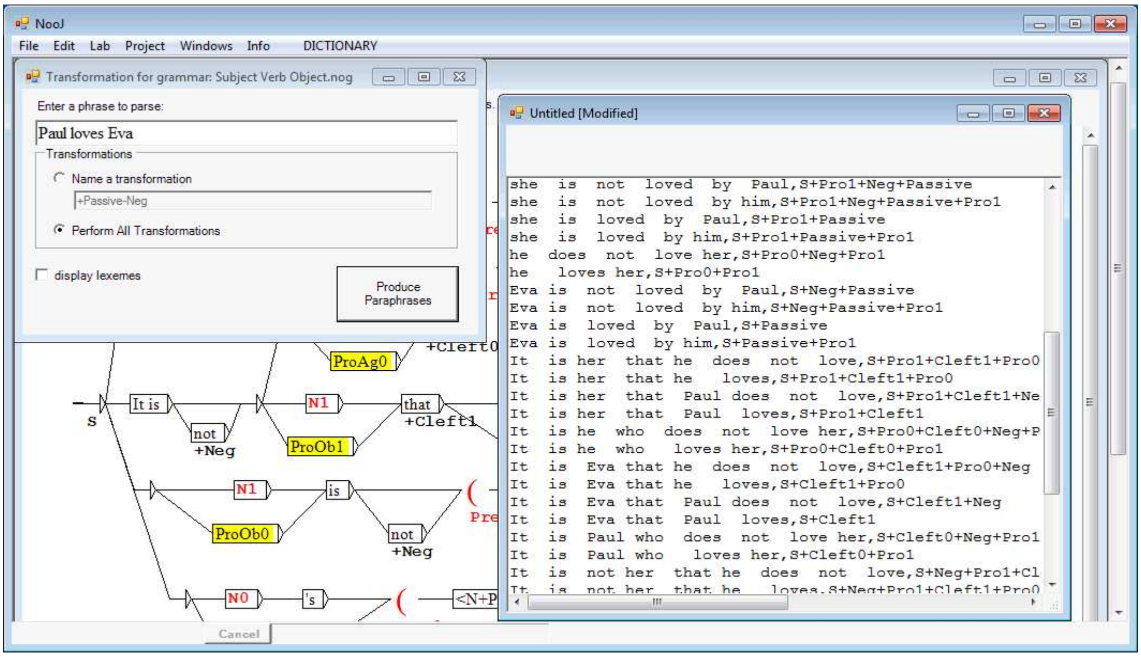


English (United States)/English (United States) syntactic grammar consists of 20 graphs.

4. Syntactic Analysis (Slim Mesfar, ENSI Tunis)

Syntactic parsing; structural *vs* derivational syntactic trees.

How to produce paraphrases automatically.



Enter a phrase to parse:

Paul loves Eva

Transformations

○ Name a transformation

+Passive-Neg

● Perform All Transformations

□ display lexemes

```
she  is  not  loved  by  Paul,S+Pro1+Neg+Passive
she  is  not  loved  by him,S+Pro1+Neg+Passive+Pro1
she  is  loved  by  Paul,S+Pro1+Passive
she  is  loved  by him,S+Pro1+Passive+Pro1
he  does  not  love her,S+Pro0+Neg+Pro1
he   loves her,S+Pro0+Pro1
Eva is  not  loved  by  Paul,S+Neg+Passive
Eva is  not  loved  by him,S+Neg+Passive+Pro1
Eva is  loved  by  Paul,S+Passive
Eva is  loved  by him,S+Passive+Pro1
It  is her  that he  does  not  love,S+Pro1+Cleft1+Pro0
It  is her  that he   loves,S+Pro1+Cleft1+Pro0
It  is her  that  Paul does  not  love,S+Pro1+Cleft1+Ne
It  is her  that  Paul  loves,S+Pro1+Cleft1
It  is he  who  does  not  love her,S+Pro0+Cleft0+Neg+P
It  is he  who   loves her,S+Pro0+Cleft0+Pro1
It  is  Eva that he  does  not  love,S+Cleft1+Pro0+Neg
It  is  Eva that he   loves,S+Cleft1+Pro0
It  is  Eva that  Paul does  not  love,S+Cleft1+Neg
It  is  Eva that  Paul  loves,S+Cleft1
It  is  Paul who  does  not  love her,S+Cleft0+Neg+Pro1
It  is  Paul who   loves her,S+Cleft0+Pro1
It  is  not her  that he  does  not  love,S+Neg+Pro1+Cl
It  is  not her  that he   loves,S+Neg+Pro1+Cleft1+Pro0
```

5. Conclusion : NooJ in the CESAR Project (Tamas Varadi, Budapest Academy of Sciences)

The CESAR project (www.cesar-project.net) is an EU funded project launched in early 2011 which together with three other projects form META-NET (www.meta-net.eu). One of the central missions of META-NET is to develop an open linguistic infrastructure within the META-SHARE facility. NooJ is one of the flagship tools that the CESAR project contributes to META-SHARE. The NooJ community (www.nooj4nlp.net) is deeply embedded in META-NET. The current overhaul of NooJ is taking place as a collaborative effort within the CESAR project. The talk will briefly outline the CESAR project and what effect it may have on the NooJ community.

References

Vučković Kristina, Bekavac Božo, Silberztein Max Eds. 2012 to appear. Selected Papers from the 2011 International NooJ Conference. (22 articles). Cambridge Scholars Publishing: Newcastle.

Gavriilidou Zoe, Chatzipapa Elina, Papadopoulou Lena, Silberztein Max Eds. 2011. Selected papers from the NooJ 2010 International Conference and Workshop. (21 articles). Univ. of Thrace Ed, Greece.

Ben Hamadou Abdelmajid, Mesfar Slim, Silberztein Max Eds 2010. NooJ 2009 International Conference and Workshop. In Finite-State Language Engineering. Centre de Publication Universitaire : Sfax Tunisia.

Silberztein, Max. 2003. NooJ manual. available at the WEB site http://www.nooj4nlp.net (200 pages).