X. Blanco

USING NOOJ FOR MULTIPURPOSE ANALYSIS OF ROMANCE LANGUAGES CORPORA

1. Introduction

This paper briefly presents the freeware NooJ, conceived and developed by Max Silberztein (University of Franche-Comté), and the available resources for Romance languages, namely Catalan, French, Italian, Portuguese, Spanish and... Latin.

The interested reader will find up-to-date information and a complete documentation about the software in **http://nooj4nlp.net**, including a detailed over 200 pages NooJ Manual in the form of a downloadable *.pdf file. Concerning the linguistics resources for Romance languages, we specify below the names of the author and/or contact person for each language module. The present paper does not provide new information about NooJ or any of its modules, it is basically intended as a convenient reminder for the reader of the *Corpora2008* Proceedings about how to get access to the mentioned documentation.

2. Generalities about NooJ

NooJ is a linguistic environment that allows users to elaborate large-coverage descriptions of languages and apply them to corpora. It was created in 2002 by Max Silberztein who had a long developer's experience as author of the linguistic system INTEX.

NooJ provides tools to describe spelling variants, inflectional and derivational morphology, several kinds of lexicons (simple words, compound/complex words, frozen and semi-frozen expressions, etc.), syntax phenomena and semantic phenomena as well. It is being used, among other applications, in Advanced Information Retrieval and in Machine-Assisted Translation.

NooJ can handle sets of thousands of text files, making it ideal for dealing with large and complex corpora. It can import over 100 file formats and works with any alphabet (Arabic, Chinese, Cyrillic...).

NooJ engine uses a Text Annotation Structure (a series of pairs *position, information*) that are always synchronized with the original text file, so that the text is not modified. Linguistic annotations can be associated to simple words, compound words or discontinuous expressions. Annotated texts can be export as XML documents (and, inversely, XML documents can be parsed and imported in order to obtain NooJ annotated texts).

The computational devices included by NooJ are:

– Finite-State Transducers (FST) that associate an input, matching with a text sequence, with and output which constitutes the analysis result;
– Finite-State Automata (FSA), a particular case of Finite-State Transducers that do not have any output;
– Recursive Transition Networks (RTN), typically graphs that contain more than one graph and references to other embedded graphs, – usually used to create libraries of graphs that can be re-used in several grammars;
– Enhanced Recursive Transition Networks (ERTN), i.e. RTN that contain variables storing some parts of the matching sequences in order to be re-used later to perform some operation;
– Context-Free Grammars (CFG), that have the power of expression of a type-2 Grammar (or a non-deterministic pushdown automaton).

In some dialogue frames, like *Locate Pattern* of the TEXT menu, the user can write directly Regular Expressions (PERL Regular Expressions or NooJ Regular Expressions), that are a quick way to perform queries with the power of a type-3 grammar without having to create any specific grammar file.

3. Romance Languages Modules

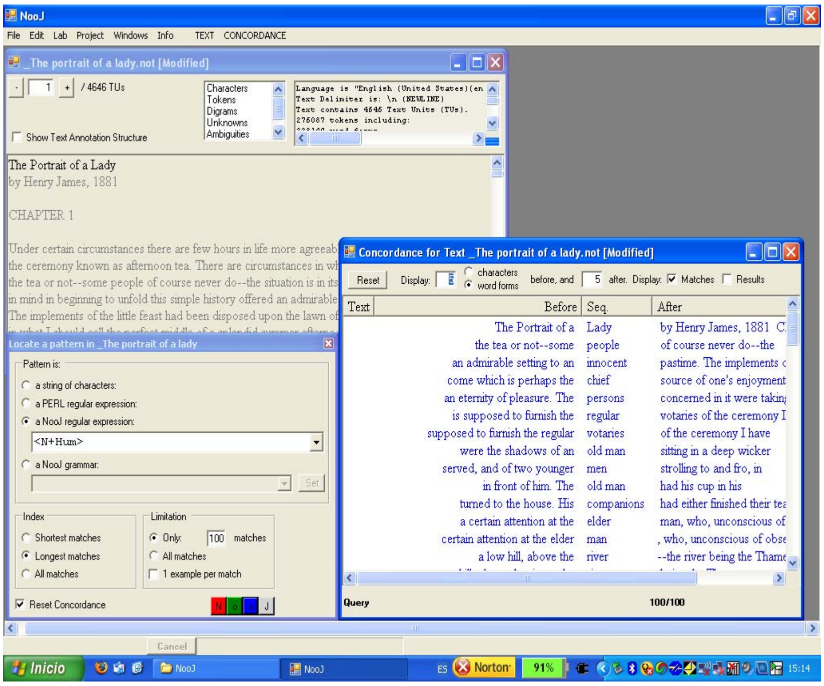
In this moment, the NooJ Community has five Romance Languages modules and a module devoted to Latin. All modules include, at least, a simple words dictionary and a sample text.

We find, in alphabetical order:

- the Catalan module, elaborated by Judith Sastre (Autonomous University of Barcelona). It includes, as corpora, a text from Mercè Rodoreda and the Universal Declaration of the Rights of Man in Catalan. It also contains fourteen morphological grammars, two examples of inflectional grammars and four lexical grammars for numerical determiners and possessive determiners and pronouns. It features a small corpus of sentences with determiners and pronouns to test some disambiguation grammars.
- the French module (Max Silberztein, University of Franche-Comté), that includes large-coverage dictionaries for simple and graphically complex words, a dictionary of first names and toponyms (cities and countries), a set of morphological grammars, examples for the inflection of simple and compound words and derivation of simple words and, finally, a set of disambiguation grammars.
- the Italian module (Simona Vietri, University of Salerno), with two sample texts, eight dictionaries, associated with their corresponding inflectional description, and a morphological grammar.
- the Portuguese module (Anabela Barreiro, University of Porto), that contains two texts, the *Declaração Universal dos Direitos Humanos* and *Viagens na Minha Terra* by Almeida Garret. It features a morphological grammar to process contracted word forms and a syntactic grammar that recognizes dates.
- The Spanish module (Xavier Blanco, Autonomous University of Barcelona), that contains a fragment of *Fortunata y Jacinta* by Benito Pérez Galdós. By request, dictionaries of proper nouns and medical terminology are available.

There is also a Latin module available (Jose Paulo Tavares, *Centro de Estudos em Letras*, Portugal) with *De bello gallico*, by C. Julius Caesar and several morphological and syntactic grammars, for instance a clitics grammar, a grammar that recognizes the verbal forms on the passive voice of the «perfectum», etc.

4. Example: Concordance <*N+Hum*>



_The portrait of a lady.not [Modified]

1 / 4646 TUs

□ Show Text Annotation Structure

```
Language is "English (United States)(en
Text Delimiter is: \n (NEWLINE)
Text contains 4646 Text Units (TUs).
276087 tokens including:
```

The Portrait of a Lady
by Henry James, 1881

CHAPTER 1

Under certain circumstances there are few hours in life more agreeab
the ceremony known as afternoon tea. There are circumstances in wh
the tea or not--some people of course never do--the situation is in its
in mind in beginning to unfold this simple history offered an admirable
The implements of the little feast had been disposed upon the lawn of

Locate a pattern in _The portrait of a lady

Pattern is:
○ a string of characters:
○ a PERL regular expression:
● a NooJ regular expression: <N+Hum>
○ a NooJ grammar:

Index: ○ Shortest matches ● Longest matches ○ All matches

Limitation: ● Only: 100 matches ○ All matches □ 1 example per match

☑ Reset Concordance

Concordance for Text _The portrait of a lady.not [Modified]

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

Text	Before	Seq.	After
	The Portrait of a	Lady	by Henry James, 1881 C
	the tea or not--some	people	of course never do--the
	an admirable setting to an	innocent	pastime. The implements c
	come which is perhaps the	chief	source of one's enjoyment
	an eternity of pleasure. The	persons	concerned in it were takin
	is supposed to furnish the	regular	votaries of the ceremony I
	supposed to furnish the regular	votaries	of the ceremony I have
	were the shadows of an	old man	sitting in a deep wicker
	served, and of two younger	men	strolling to and fro, in
	in front of him. The	old man	had his cup in his
	turned to the house. His	companions	had either finished their tea
	a certain attention at the	elder	man, who, unconscious of
	certain attention at the elder	man	, who, unconscious of obse
	a low hill, above the	river	--the river being the Thame

Query 100/100

Fig. 1. Screenshot: Concordance <*N+Hum*>

5. Conclusion

By way of conclusion, we would like to insist in the fact that NooJ is being continuously developed and supported not only by his author but also by a large international community that is building large-coverage linguistic descriptions for more than a dozen languages. There is an annual series of NooJ International Conferences, the last one having taken place in the Hungarian Academy of Sciences, Budapest, June 8–10, 2008. Moreover, other workshops and tutorials are regularly organized in different European countries.

It is important to emphasize that NooJ is being used as an effective pedagogical device for Computational Linguistics teaching. For example, we are using it intensively in the Autonomous University of Barcelona in the Master in *Information Processing and Multilingual Communication* and in the Doctoral Program in *Romance Languages and Cultures* as well.

A book including about twenty papers with exercises and pedagogical suggestions for selected grammatical issues in different languages is being prepared by the NooJ Community (X. Blanco and M. Silberztein, forthcoming). Each contribution will be associated with downloadable NooJ linguistic data. The languages treated will include: Albanese, Arabic, Armenian, Bulgarian, Catalan, Chinese, English, French, Greek, Hebrew, Hungarian, Italian, Korean, Latin, Polish, Portuguese, Rumanian, Serbian and Spanish.

6. Bibliography

Here are relevant publications on NooJ.

Blanco X.; Silberztein M. Proceedings of the 2007 International NooJ Conference, Cambridge Scholars Publishing. [forthcoming]

Blanco X.; Silberztein M. Processing Natural Languages with NooJ. Servei de Publicacions de la Universitat Autònoma de Barcelona. [forthcoming]

Koeva S.; Maurel D.; Silberztein M. Formaliser les langues avec l'ordinateur: de INTEX à NooJ. Presses Universitaires de Franche-Comté, 2007.

Muller R.; Silberztein M. (eds.) INTEX pour la Linguistique et le traitement automatique des langues. Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté, 2004.

Silberztein M. NooJ Manual. 2006 // URL: **http://nooj4nlp.net**