Construction of linguistic resources for the extraction of complex text segments

PARSEME (PARSing and Multi-word Expressions)
Working Group 2

T. Kyriacopoulou*, C. Martineau*, C. Martinez*, A. Fotopoulou*

*LIGM, Université Paris-Est Marne-La-Vallée, France.
ILSP, “Athena” RIC Greece.

Athens, 10 March 2014
Objective

Develop reusable electronic linguistic resources to identify, annotate and normalize complex text segments in French and Modern Greek.

Approach

1. Design and build dictionary graphs to identify and normalize specific entities.
2. Use traditional DELA dictionaries, dictionary graphs and local grammars together in order to formulate complex pattern queries.

Simple local grammar example

This grammar identifies segments such as:
- lundi 10 mars 2014
- vendredi 13 décembre 2006
But isn’t designed to handle simple queries such as:
- only dates without a year number.
- dates with at least a day name.

Using ordinary grammars it isn’t possible to exercise a fine-grained control over queries (e.g. employing grammatical, semantic or inflectional constraints).
Dictionary graph (toy example)

Dictionary graphs dynamically produce new text dictionary entries as normal DELA-lines, including grammatical, semantic and inflectional rules.

Now is possible to use lexical masks like `<Date~a>` and `<Date+nj>`

Identifying relevant complex text segments

Using dictionary attributes it's possible to exercise a more fine-grained identification of each complex text segment.

This graph looks for a person name (`Person`), composed by a given name (`+prenom`) and a family name (`+patronyme`), followed by a verbal form (`FV`) in compound past tense (`+K:P`) without negation (`~Neg`), and by a date (`Date`) with at least a day (`+j`) and a month (`+m`) numbers. e.g. this construction recognizes sentences like:

"Yves Delanoue est finalement arrivé mardi 11 mars"
(trans.: Yves Delanoue finally arrived Tuesday, March 11)