

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



LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE



Ερευνητικό Κέντρο Αθηνών
Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες
της Πληροφορίας, των Επικοινωνιών, της Γλώσσας

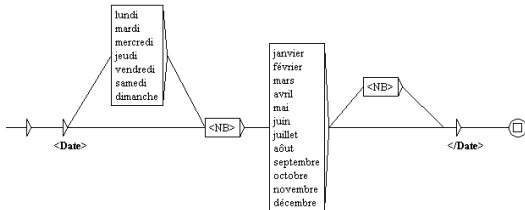
Objective

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*

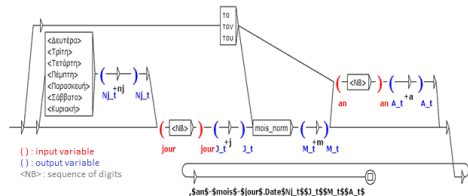
But isn't designed to handle simple queries such as:

- *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.



```

Πέμπτη 31 Ιουλίου, -07-31.Date+nj+j+m
Πέμπτη 4 Δεκεμβρίου, -12-4.Date+nj+j+m
Παρασκευή 8 Αυγούστου, -08-8.Date+nj+j+m
Παρασκευή 11 Ιουλίου, -07-11.Date+nj+j+m
ΕΣΑΒΒΑΤΟ 30 Αυγούστου, -08-30.Date+nj+j+m
ΕΣΑΒΒΑΤΟ 1 Νοεμβρίου, -11-1.Date+nj+j+m
Εσάββατο 2 Αυγούστου, -08-2.Date+nj+j+m
ΤΟ 1987,1987--.Date+a
ΤΟΝ ΜΑΡΤΙΟ, -05-.Date+m
ΤΟΝ φεβρουάριο, -02-.Date+m
ΤΟΥ 2004,2004--.Date+a
Τετάρτη 15 Οκτωβρίου, -10-15.Date+nj+j+m
Τετάρτη 2 Ιουλίου, -07-2.Date+nj+j+m
Τετάρτη 29 Ιουλίου, -07-29.Date+nj+j+m
Τετάρτη 30 Απριλίου, -04-30.Date+nj+j+m
Τετάρτη 5 Ιουλίου 1961,1961-07-5.Date+nj+j+m+a
    
```

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)