LABORATOIRE D'INFORMATIQUE GASPARD-MONGE

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

Construction of linguistic resources for the extraction of «complex text» segments

Tita Kyriacopoulou[1], Claude Martineau[1], Cristian Martinez[1], Aggeliki Fotopoulou[2]

[1]LIGM, Université Paris-Est Marne-La-Vallée, France — {tita,claude.martineau,cristian.martinez}@univ-paris-est.fr

[2]ILSP, "Athena" RIC Greece — afotop@ilsp.athena-innovation.gr

Ερευνητικό Κέντρο Αθηνά
Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών, της Γνώσης

WG2

- **Objective : Fine-grained annotation of complex text segments in French and Modern Greek including:**
  - Compound-words, e.g. *πιστωτική κάρτα* (*credit card*);
  - Entity names, e.g. *Pierre Durand*, *10 mai 2014, 1700 euros par mois* (*per month*);
  - Verbal forms: complex forms of verbs with compound tense, negation, embedded modifiers, e.g.
    - *Max n'est pas encore arrivé* (*Max has not yet arrived*);
  - Frozen expressions, e.g. *Faire d'une pierre deux coups* (*to kill two birds with one stone*)

- **Tools : Unitex [1], an open source, cross-platform and multilingual corpus processing suite which supports:**
  - **DELA (Dictionnaires Électroniques du LADL / LADL electronic dictionaries)**

    A typical DELA entry is composed by a simple or compound inflected form, followed by a lemma and grammatical information, each entry can be associated with syntactic and semantic attributes and inflection rules:

    ```
    inflected_form,lemma.grammatical_information+attributes:inflection_rule
    ```

    Take for example the French compound word "*bateau amiral*" (flagship), a DELA representation could be:

    ```
    bateau amiral,.N+NA+Conc+z3:ms
    bateaux amiraux,bateau amiral.N+NA+Conc+z3:mp
    ```

  - **Rules called «local grammars» represented by graphs**
    - Graphical representations of local grammars are composed by a set of linked boxes.
    - A legal path is a path between initial and final states.
    - For instance, the graph of Fig. 1. recognizes French date-entities.

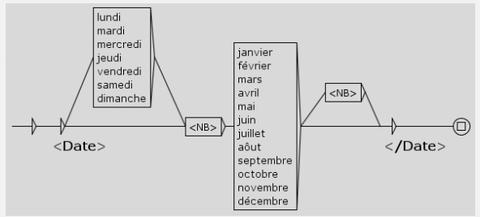


Fig.1. Date recognition in French

- **Benefits & Limits of simple/ordinary grammars**
  - ✓ Grammars can identify sequences of words that would be **impossible** to include in an electronic dictionary.
  - ✓ **However**, if we need, for instance, to express the fact that date-entities could be sometimes identified only when the name day or the year number are not provided, it would be compulsory to build a new set of graph versions, which turns out to be a very costly process for the development and maintenance of grammar resources.

- **Dictionary graphs :** Recent Unitex improvements allow to design *dictionary graphs.* Graphs' outputs which produce **dynamically** new text dictionary entries as normal DELA lines, including the construction of syntactic and semantic attributes and inflection rules. In addition to sequences previously identified by the graph in Fig.1., the graph in Fig. 2. distinguish the forms "`en month_name`" (e.g. *en avril*) and "`en year_number`" (e.g. *en 2014*).

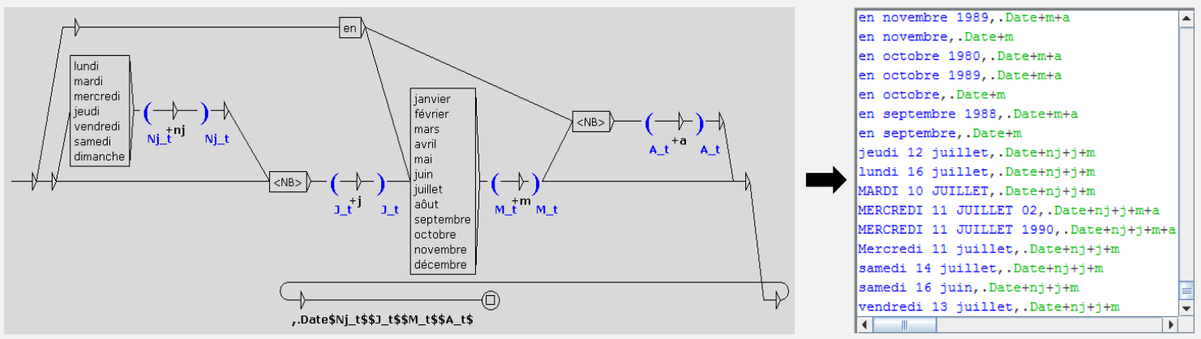


Fig. 2. Building date-entities dictionary entries in French

- **Using new generated entries**
  - Graph in Fig.2. produces new dictionary entries having *Date* as grammatical category and including a set of attributes to indicate the presence or absence of a portion of the date (`+nj` : day name, `+j` : day number, `+m` : month name, or `+a` : year number).
  - It is then possible to exercise more fine-grained control in producing graphs which are looking for date named-entities :
    - For example, the lexical mask ***<Date+j+m+a>*** selects only full date expressions.
    - If we are interested in dates where only day number and month name are present, we could use the lexical mask ***<Date+j+m~nj~a>,*** here the tilde grapheme (`~`) is used to exclude name day (`nj`) and year number (`a`) codes.

- **Building a reference form**
  - The canonical form (`lemma`) in a common DELA is also considered as the reference form.
  - For a name-entity without lemma we could use as reference form a normalized version of the inflected form.
  - For example, the local grammar in Fig.3. builds a set of Modern Greek normalized DELA-entries able to identify date-entities.

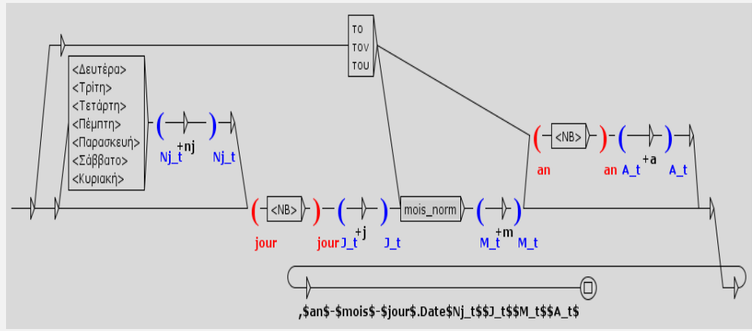


Fig. 3. Building normalized date entries in Modern Greek

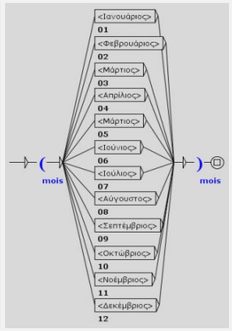
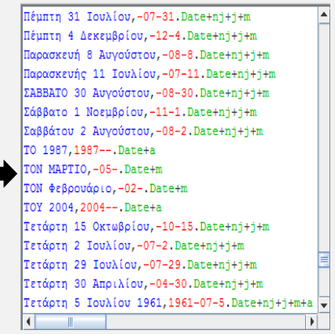


Fig. 4. Month normalisation



- **Identifying relevant complex text segments**

  Using dictionary graphs we can easily express complex pattern queries:

Fig. 5. Complex pattern query

Graph in Fig. 5. 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 a date (`Date`) with at least a day (`+j`) and a month (`+m`). e.g. this construction recognizes sentences like :

Yves Delanoue est finalement arrivé mardi 11 mars 2014 (*Yves Delanoue finally arrived Tuesday, March 11, 2014*)

- **Conclusion**
  - The identification of complex sequences of text segments using dictionary graphs, which combining the power and versatility of the local grammars and the expressivity of the electronic dictionaries, is an effective method to promote the **adaptability**, **reusability** and **modularity** of electronic linguistic resources.
  - Already used for French and Modern Greek languages, our approach will be soon extended to other languages.

[1] **http://igm.univ-mlv.fr/~unitex**