Construction of linguistic resources for the extraction of ≪ complex text segments ≫

PARSEME Working Group 2

Tita Kyriacopoulou*, Claude Martineau*, Cristian Martinez*, Aggeliki Fotopoulou†

∗ LIGM, Université Paris-Est Marne-La-Vallée, France.

{*tita,claude.martineau,cristian.martinez*}*@univ-paris-est.fr*

† ILSP, "Athena" RIC Greece.

afotop@ilsp.athena-innovation.gr

The development of computational linguistic resources (electronic dictionaries and grammars) for the automatic extraction, identification, and further fine-grained annotation of ≪ complex text segments ≫, is the core of our work. We use and extend the notion of *multi-word units* (MWUs) by allowing a large description of linguistic objects: compound nouns, entity names, verbal forms (compound tense and negate forms, introduction of clauses between the auxiliary and the past participle, etc.) and frozen expressions (i.e. idioms).

The introduced resources are builded using Unitex (Paumier, 2003), an open source, cross-platform and multilingual corpus processing suite. Unitex tools are designed to perform several natural language processing (NLP) tasks on a textual corpus relying on linguistic resources such as electronic dictionaries and grammars represented as *finite state transducteurs* (FSTs), *recursive transition networks* (RTNs) (Woods, 1970) and lexicon-grammars (Gross, 1996).

Unitex supports the DELA (*Dictionnaires Électroniques du LADL* : LADL electronic dictionaries) standard of lexicon formats and apply technologies designed at LADL (*Laboratoire d'informatique documentaire et linguistique*, University of Paris 7, France). A typical DELA entry, shown below, 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
```

Graphical representations of local grammars, called *graphs* under Unitex, are composed by a set of linked boxes of different kinds (initial, final, transduction, record start/stop and comment boxes), graph outputs can be displayed as an interactive concordancer or even can be added or substituted in the recognized text segments. Local grammars are used, as shown as example in the next figure, to identify and annotate sequences of words representing French date named-entities. Notice that <NB> is an Unitex special lexical mask designed to match any contiguous sequence of digits.

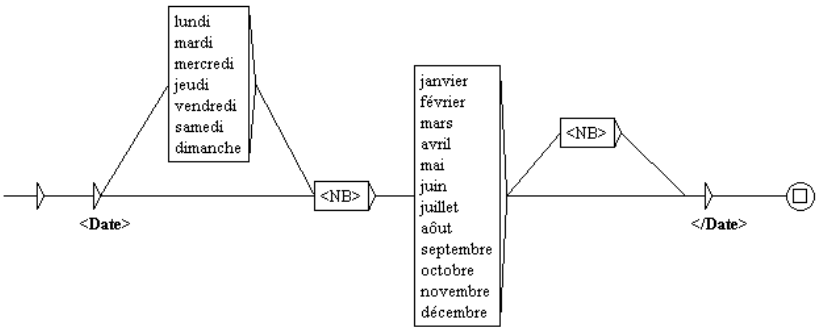

Figure 1. *Graph that recognizes date named-entities*

Clearly, the grammar in Figure 1, can identify sequences of words that would be impossible to include in an electronic dictionary. However, if we need for instance express the fact that a date segment could be sometimes identified only when the day name 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 expensive process for the development and maintenance of *named entity recognition* (NER) systems.

Starting with version 2.0 of Unitex, it is possible to design dictionary graphs, the output of such graphs produce new text dictionaries entries as normal DELA-lines, these include the construction of syntactic and semantic attributes and inflection rules. Dictionary graphs have exactly the same properties than normal graphs and can use results given by previous dictionaries.

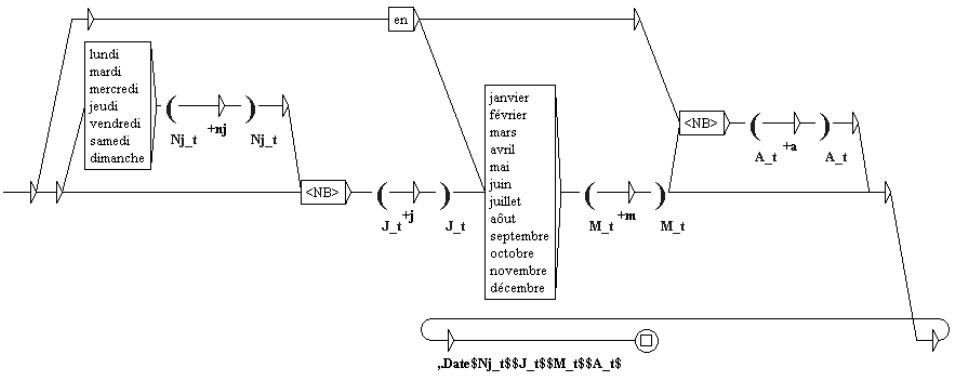

Figure 2. *Dictionary graph that recognizes date named-entities*

The graph shown above was originally based on the grammar presented in Figure 1. In addition to sequences previously identified, this graph distinguish the forms `"en month_name"` (e.g. *en avril*) and `"en year_number"` (e.g. *en 2014*). The last graph output: `,.Date$Nj_t$$Jt_t$$M_t$$A_t$` is in charge to dynamically produce new dictionaries 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). In this way, dates such *"25 avril 2000", "en janvier", "lundi 3 mars", "en 2014"*, automatically produce respective dictionary entries as follows:

```
25 avril 2000,.Date+j+m+a    en janvier,.Date+m
lundi 3 mars,.Date+nj+j+m    en 2014,.Date+a
```

Using the generated entries it is now possible to exercise more fine-grained control when producing graphs that looking for date named-entities. Take as an example the lexical mask <`Date+j+m+a`> which selects only full date expressions, note that the presence of the day name is optional. In another scenario, when 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 excludes day name (`nj`) and year number (`a`) codes. More sophisticated graphs can be design to perform tasks such as date normalization (e.g. *25 avril 2000* ⇒ *25/04/00*) or date-entities context processing (e.g. *"**vers** le 10 mai **dernier**"*, here the words in bold represent the context).

These dictionary graphs can be manually elaborated in the same way that they are constructed in the person name recognition system presented in Kyriacopoulou et al. (2011), furthermore, they can be automatically generated from other types of pre-existing text-based resources. This possibility has been explored in a previous work of Tolone et al. (2012) focused on the generation of french composed adverbs from Lexicon-Grammar tables (Gross, 1975; Fotopoulou, 1993) which are represented using the LGLex syntactic lexicon format, the dictionary graph generated is able to produce adverb entries (Text between brackets [ ] indicates an alternative segment) such as the following:

```
par temps [couvert],.ADV+pca          à [trois] francs près,.ADV
```

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 seems to be an effective method to promote the adaptability, reusability and modularity of linguistic resources. Already used for French and Modern Greek languages our approach will be soon extended to other languages.

References

[1] A. Fotopoulou. *Une classification des phrases à compléments figés en grec moderne: étude morphosyntaxique des phrases figées*. 1993.

[2] M. Gross. "Lexicon grammar". In: *Concise Encyclopedia of Syntactic Theories*. Ed. by Keith Brown and Jim Miller. Oxford: Elsevier Science, 1996, pp. 244–258.

[3] M. Gross. *Méthodes en syntaxe: régime des constructions complétives*. Actualités scientifiques et industrielles. Hermann, 1975.

[4] Tita Kyriacopoulou, Claude Martineau, and Thanassis Mavropoulos. "Les noms propres de personne en français et en grec : reconnaissance, extraction et enrichissement de dictionnaire". In: *Proceedings of the 30th Conference on Lexis and Grammar*. Nicosie, Chypre, 2011, p. 8.

[5] S. Paumier. "Unitex 3.1beta : User Manual". In: *Université Marne-la-Vallée* (2003).

[6] Elsa Tolone et al. "Extending the adverbial coverage of a French morphological lexicon". In: *Proceedings of poster session of the the 8th Language Resources and Evaluation Conference (LREC'12)*. Istanbul, Turkey, May 2012.

[7] W. A. Woods. "Transition Network Grammars for Natural Language Analysis". In: *Commun. ACM* 13.10 (Oct. 1970), pp. 591–606.