Using Lexicon-Grammar tables for French verbs in a large-coverage parser

Elsa Tolone[1] & Benoît Sagot[2]

1. IGM, Université Paris-Est (France)
2. Alpage, INRIA Paris-Rocquencourt & Université Paris 7 (France)

4th Language and Technology Conference – Poznań, Poland
November 7, 2009

Context

- **Lexicon-Grammar tables** are a large-coverage lexical resource developed manually for years
- They contain **syntactic** and semantico-syntactic information
- Such information is arguably very **useful for parsing**
- But Lexicon-Grammar tables are **not usable as such** in a parser
  - features that are shared by all entries in a given table are not explicitely given
  - lexical features are not properly formalized
  - these data need to be integrated in a real-life parser

Objectives

- Three major objectives
  1. **convert** Lexicon-Grammar tables to an NLP format,
  2. plug the resulting lexicon, named *lglex*$_{\mathrm{Le}\mathit{fff}}$, with a **parser**
  3. **evaluate** the resulting parser
- NLP tools used:
  - parser: FRMG [Thomasset et de La Clergerie 2005]
  - lexical formalism: Alexina, formalism used by the Le*fff* lexicon [Sagot *et al.* 2006] used by FRMG

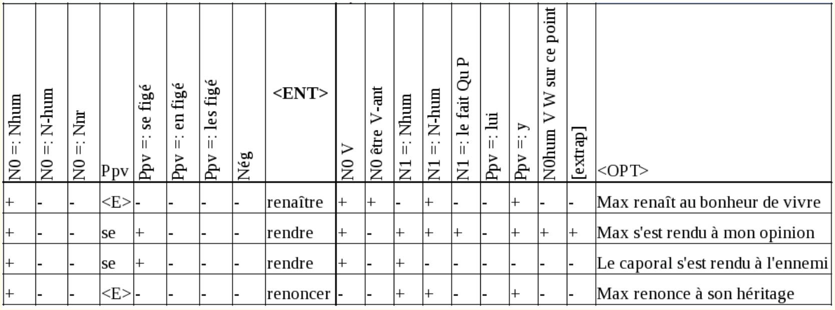
  → this allows for a comparison between FRMG$_{\mathrm{Le}\mathit{fff}}$ and FRMG$_{\mathit{lglex}}$
- In this work, we considered only **simple verbs**

1. Lexicon-Grammar verb tables for French

Lexicon-Grammar tables

- a verb class is defined by a set of **“defining features”**
- for a given table, the defining features often include:
  - a basic defining feature, often a subcategorization frame,
  - often additional features (distributional, morphological, transformational, semantic,etc.)
- These features define **61 verb classes**
- Each verb class is described in a **table**:
  - one row for each (lemma-level) entry
  - one column for each feature that is relevant for the class
  - at the intersection of a row and a column, $+$ (resp. $-$) $=$ the corresponding feature is valid (resp. not valid) for the corresponding entry

Table V33



N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	**<ENT>**	N0 V	N0 être V-ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<OPT>
+	-	-	<E>	-	-	-	-	renaître	+	+	-	+	-	-	+	-	-	Max renaît au bonheur de vivre
+	-	-	se	+	-	-	-	rendre	+	-	+	+	+	-	+	+	+	Max s'est rendu à mon opinion
+	-	-	se	+	-	-	-	rendre	+	-	+	-	-	-	-	-	-	Le caporal s'est rendu à l'ennemi
+	-	-	<E>	-	-	-	-	renoncer	-	-	+	+	-	-	+	-	-	Max renonce à son héritage

Defining feature: N_0 V *à* N_1

Table of classes

Defining features are not represented in the tables
→ to be dealt with in a **table of classes** for simple verbs:

- one row for each class
- one column for each feature (overall, after normalization among tables)
- at the intersection of a row and a column,
  - o = the feature is explicitely dealt with in the corresponding table
  - $+$ (resp. $-$) = the corresponding feature is valid (resp. not valid) for all entries in the corresponding class

The table of simple verb classes has just been completed [Constant & Tolone 2008]

The table of simple verb classes

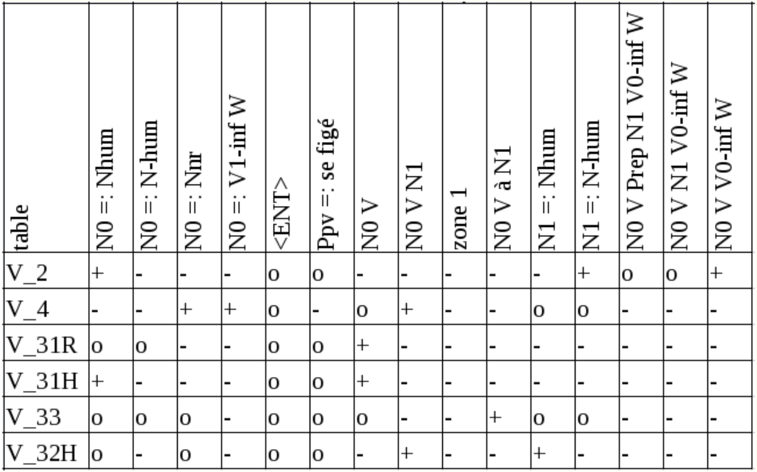


table	N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	N0 =: V1-inf W	<ENT>	Ppv =: se figé	N0 V	N0 V N1	zone 1	N0 V à N1	N1 =: Nhum	N1 =: N-hum	N0 V Prep N1 V0-inf W	N0 V N1 V0-inf W	N0 V V0-inf W
V_2	+	-	-	-	o	o	-	-	-	-	-	+	o	o	+
V_4	-	-	+	+	o	-	o	+	-	-	o	o	-	-	-
V_31R	o	o	-	-	o	o	+	-	-	-	-	-	-	-	-
V_31H	+	-	-	-	o	o	+	-	-	-	-	-	-	-	-
V_33	o	o	o	-	o	o	o	-	-	+	o	o	-	-	-
V_32H	o	-	o	-	o	o	-	+	-	-	+	-	-	-	-

lglex

The table of simple verb classes permits the extraction of a **syntactic lexicon** of simple verbs from Lexicon-Grammar tables [Constant & Tolone 2008]:

- text or XML format
- named **lglex**
- generated from the original Excel verb tables by the *LGExtract* tool

lglex is the starting point of the conversion process towards the Alexina format

lglex: an example

```
ID=V_35L_242
lexical-info=[locs=(loc=[id="1",list=()],loc=[id="2",list=()]),cat="verb",verb=[lemma="ruisseler"],
    aux-list=(),prepositions=()]
args=(
    const=[dist=(comp=[cat="NP",source="true",introd-prep=(),origine=(orig="Loc N1 =: de N1 source"),
        introd-loc=(prep="de")]),pos="1"],
    const=[dist=(comp=[ cat="NP",introd-prep=(),origine=(orig="Loc N2 =: vers N2 destination",
        orig="Loc N2 =: dans N2 destination"),introd-loc=(prep="vers",prep="dans"),destination="true"]),pos="2"],
    const=[pos="0",dist=(comp=[cat="NP",introd-prep=(),nothum="true",origine=(orig="N0 =: N-hum"),
        introd-loc=()])])
all-constructions=[absolute=(construction="o::N0 V Loc N1 source Loc N2 destination",construction="o::N0 V",
    construction="o::N0 être V-ant",construction="true::N0 V Loc N1"),
    relative=(construction="Ppv =: y",construction="Ppv =: en",construction="[extrap]")]
example=[example="L'eau ruisselle de la gouttière sur les passants"]
```

2. The Le*fff* and the Alexina format

Le Le*fff*

- The Le*fff* (Lexique des Formes Fléchies du Français) is a morphological and syntactic lexicon for French
  - large coverage (536,375 entries corresponding to 110,477 distinct lemmas covering all categories)
  - freely available (LGPL-LR license)
- It relies on the **Alexina** framework for the modeling and acquisition of morphological and syntactic lexicons.

Alexina

Two-level architecture

- The **intensional** lexicon
  - associates with each entry (meaning of a lemma) a canonical subcategorization frame
  - lists all possible redistributions (restructurations) from this frame
- The **compilation** process of the intensional lexicon into the **extensional** lexicon generates different entries for each inflected form and each possible redistribution.

Alexina on an example

- Example of an intensional entry:

  clarifier$_1$ *v-er:std*
  Lemma;v;
  <**Suj**:*cln|scompl|sinf|sn*,**Obj**:*(cla|scompl|sn)*>;
  %active,%se_moyen_impersonal,
  %passive_impersonal,%passive

3. Converting *lglex* into an Alexina lexicon

Overview of the conversion process

- The conversion of Lexicon-Grammar tables into the Alexina framework is **not straightforward**
  - It requires a **formal definition** or a **dynamic interpretation** of all feature names
  - Directly or indirectly, these features may:
    - specify full subcategorization frames
    - specify partial information about subcategorization frames (the fact that an argument is not mandatory, a possible realization of an argument, etc.)
    - correspond to a redistribution
    - lead to the construction of an additional entry
  - Additional important information must be gathered heuristically or from other lexical resources
    - the name of each syntactic function, attribution phenomena, morphological information, etc.
- We won't enter into the details of this conversion process.

The previous example after conversion

ruisseler$^{35\mathrm{L}}_{242}$ *v-er:std*
100;Lemma;v;
*<***Suj***:cln|sn,***Dloc***:(de-sn|en),***Loc***:(vers-sn|dans-sn|y)>;*
cat=v;
%active

The resulting lexicon: *lglex*$_{\text{Le}\textit{fff}}$

The resulting verb lexicon, *lglex*$_{\text{Le}\textit{fff}}$, contains 16 903 entries for 5 694 unique verb lemmas (2,96 entries per lemma).

- to be compared with the last published version of the Le*fff*: 7 072 verb entries for 6 818 unique verb lemmas (1,04 entries per lemma)

At the extensional level, the Le*fff* contains 361 268 entries, whereas *lglex*$_{\text{Le}\textit{fff}}$ contains 763 555 entries.

4. Integration in the FRMG parser

Integration in the FRMG parser

- We replaced the Le*fff* with a modified version of the Le*fff* in which verb entries are replaced by *lglex*$_{\mathrm{Le}\mathit{fff}}$
- additional Le*fff* entries must be added for
  - (semi-)auxiliaries
  - several raising verbs
  - impersonal verb constructions
  - light verbs

The result is a **variant of FRMG**, named FRMG$_{\mathit{lglex}}$ unlike the standard variant denoted by FRMG$_{\mathrm{Le}\mathit{fff}}$.

5. Evaluation and discussion

Protocol used

- We evaluated FRMG$_{\mathrm{Le}\textit{fff}}$ and FRMG$_{\textit{lglex}}$ by parsing the manually annotated part of the EASy corpus [Paroubek *et al.* 2005]
  - 4 306 sentences of various genres (journalistic, medical, oral, questions, literacy, etc.)
- evaluation metrics: those of the first EASy parsers' evaluation campaign that took place in December 2005 [Paroubek *et al.* 2006]
  - evaluation in **chunks** and **relations** (∼ dependencies between lexical words)

Preliminary remarks

FRMG_{lglex}'s results must be analyzed with the following facts in mind:

- FRMG_{lglex}'s verb entries are the result of a conversion process from the original tables
  → this conversion process certainly introduces errors
- the Le*fff* was developed in parallel with the EASy campaigns (unlike Lexicon-Grammar tables)

Results

Comparative results of FRMG$_{\text{Lefff}}$ and FRMG$_{\textit{lglex}}$ (in terms of f-measure):

	Chunks		Relations	
Sub-corpus	FRMG$_{\text{Lefff}}$	FRMG$_{\textit{lglex}}$	FRMG$_{\text{Lefff}}$	FRMG$_{\textit{lglex}}$
general_lemonde	**86.8%**	82.8%	**59.8%**	56.9%
general_senat	82.7%	**83.1%**	**56.7%**	54.9%
litteraire_2	**84.7%**	81.5%	**59.2%**	56.3%
medical_2	85.4%	**89.2%**	**62.4%**	58.6%
oral_delic_8	**74.1%**	73.6%	47.2%	**48.5%**
questions_amaryllis	90.5%	**90.6%**	**65.6%**	63.2%
total	**84.4%**	*82.3%*	**59.9%**	*56.6%*

Parsing times higher with FRMG$_{\textit{lglex}}$ than with FRMG$_{\text{Lefff}}$: the median parsing time per sentence is 0,62s vs. 0,26s

- this comes from the higher average number of entries per verb lemma (approx. 3) in *lglex* than in the Lefff

- FRMG$_{\mathit{lglex}}$ gives better results than FRMG$_{\mathrm{Le}\mathit{fff}}$ for some relations
  - “standard” relations MOD-A et MOD-R
  - “tough” relations MOD-P et APP
- the ATB-SO relation (subject or object attribute) is the relation with the highest difference in terms of recall (34,0% vs. 58,4%)
  - this is because Lexicon-Grammar tables encode very little information about attribution phenomena

- the higher **lexical ambiguity** in FRMG$_{\mathit{lglex}}$ leads to
  - a higher ambiguity for the parser
  - and therefore a higher error rate in the disambiguation step
- example:
  - *[...] on estime que cette décision [ferait] dérailler le processus de paix*
    ([...] it is considered that this decision [would] make the peace process fail
  - FRMG uses the standard following heuristics: “arguments are prefered to modifiers”
  - FRMG$_{\mathit{lglex}}$ considers *de paix* as an argument of *estimer* (*estimer qqch de qqn*)
  - FRMG$_{\mathrm{Le}\mathit{fff}}$ makes no error since in the Le*fff*, *estimer* has no Objde

Conclusions et perspectives

Short-term

- Many sentences receive a full parse from FRMG$_{\textit{lglex}}$ but not from FRMG$_{\text{Le}\textit{fff}}$, and vice-versa
  - → **coupling both parser variants** could prove useful, since full parses have a higher f-measure than partial parses
- Le*fff* and *lglex*$_{\text{Le}\textit{fff}}$ are **complementary** in many aspects
- → use automatic techniques to improve each resource thanks to the other (e.g., via statistical analysis of parsing results [Sagot et de La Clergerie 2008])

Long-term

Optimize the use of lexical data in Lexicon-Grammar for parsing

- **improve/correct the conversion process**
- generalize the technique to Lexicon-Grammar tables for **other categories**
- generalize the technique to **other languages** for which large-coverage Lexicon-Grammar tables are available (e.g., Greek)