Development of a lexicon of French (Desarollo de un léxico del francés)

Elsa Tolone[1]

1. LIGM, Université Paris-Est (France)

I Jornadas de Jóvenes Lingüistas, Buenos Aires, Argentina
March 21, 2011

Context in 1970's

- **Syntactic lexicon** for NLP applications
  $\rightarrow$ syntactic and semantic information of predicates (verb, noun or adjective)
- Works in syntax: create general rules, i.e., transformation rules of Chomsky
  $\rightarrow$ the question is **For each word, this general rule can be applied ?**
- Objective of M. Gross: to create a large-coverage lexical resource
  $\rightarrow$ **Lexicon-Grammar tables**

Context now

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

1. Lexicon-Grammar tables for French

Lexicon-Grammar tables

Developed **manually** for over 40 years by [Gross 1975] and the Computational Linguistics Group of LIGM (Université Paris-Est)

- Use in French of verbs, adverbs, predicative nouns and adjectives and frozen expressions
  $\rightarrow$ they **share some features**
- Study the syntax in a basic sentence
  (or **subcategorization frame**)
  e.g.: N0 V N1
- The different meanings are distinguished
  e.g.: *se rendre* (*sb to surrender to sb/sb to accept sth*)

Principle

- Each 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
- A 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.)
    e.g.: N0 =: Nhum $\rightarrow$ names of people

Table V_33

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: N0 V à N1

Number of entries/category

- Inventory [Tolone 2009]:
  - 64 classes of **simple verbs**
    - 13 862 entries for 5 739 distinct lemmas
  - 32 classes of **simple and frozen adverbs** (adverbs in *-ment* and frozen adverbs)
    - 10 487 entries for 9 273 distinct lemmas
      e.g.: *[changer] du jour au lendemain* (*[to change] overnight*)
  - 78 classes of **simple and frozen predicative nouns** (nouns with argument(s)that are studied with their light verb)
    - 12 696 entries for 8 530 distinct lemmas
      e.g.: *Luc monte une attaque contre le fort*
      (*Luc is lauching an attack against the fort*)
  - 69 classes of **verbal frozen expressions**
    - 39 627 entries for 38 626 distinct lemmas
      e.g.: *Tu n'arrives pas à la cheville de Marie*
      (*You can't hold a candle to Mary*,
      literaly *You don't arrive at the ankle of Mary*)

Improvement of the tables

- Problems:
  - **Implicit defining features** (literature)
    → Added in the tables
  - **Different names** for the same feature
    → Harmonization of the column headings
  - Features **not defined clearly**
    → Documentation of features
- Conversion into a syntactic lexicon *LGLex*, based on the **same linguistic concepts** that tables
  → text or XML format generated from the original Excel tables by the *LGExtract* tool [Constant & Tolone 2010]
- Conversion into the Lefff format, used in a **parser**
  → is **not straightforward** because it requires a formal definition or a dynamic interpretation of all feature names [Tolone & Sagot 2011]

2. *LGLex* and *LGLex*$_{\mathrm{Lefff}}$

The Lefff lexicon

- The Lefff (Lexique des Formes Fléchies du Français) is a morphological and syntactic lexicon for French [Sagot 2010]
  - large coverage (536 375 entries corresponding to 110 477 distinct lemmas covering all categories)
  - freely available (LGPL-LR license)
- 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

LGLex: an example of verb

ID=V_33_131
*lexical-info=[cat="verb",verb=[***lemma="rendre",ppvse="true"****],*
aux-list=(etre="true"),prepositions=(),locatifs=()]
args=(
*const=[***pos="0"****,dist=(comp=[***cat="NP",hum="true"****,introd-prep=(),introd-loc=(),*
origin=(orig="N0 =: Nhum")],
*const=[***pos="1"****,dist=(comp=[***cat="NP",hum="true"****,introd-prep=(),introd-loc=(),*
origin=(orig="N1 =: Nhum")]))])
*all-constructions=[absolute=(construction="true::***N0 V à N1***",construction="o::***N0 V***",*
relative=()]
example=[example="Le caporal s'est rendu à l'ennemi"]

- entry *se rendre* V_33_131 (*to surrender*)

LGLex$_{\mathrm{Lefff}}$: the previous example after conversion

rendre V_33_131 *v-re3*
100;se Lemma;v;
<**Suj**:*cln*|*sn*,**Objà**:(*à-sn*)>;
*cat=v,@pron,***@SujNhum**;**@ObjàNhum**;
%actif,

- *Vercingetorix s'est rendu à Cesar*
  (*Vercingetorix surrendered to Ceasar*)
  - *Vercingetorix s'est rendu*
    (*Vercingetorix surrendered*)

LGLex$_{\mathrm{Lefff}}$: the previous example after conversion (2)

rendre V_33_131 *v-re3*
100;se Lemma;v;
*<***Suj***:cln|sn,***Objà***:(à-sn)>;*
*cat=v,@pron,***@SujNhum***;***@ObjàNhum***;*
%actif,

Entry in the **intensional** lexicon:

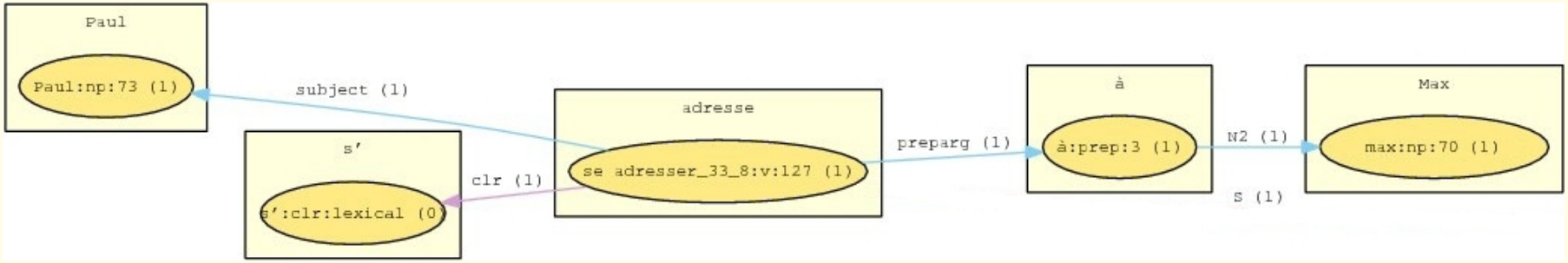
- entry identifier: categorie_numTable_numEntry
- morphological class, which defines the patterns that build its inflected forms, using inflection classes from the Lefff
- weight
- category (or part-of-speech)
- initial sub-categorization frame with syntactic functions (Suj, Obj, Objde, Loc, etc.) and the possible realizations (par-sn, sinf, de-scompl, cln, etc.) for each function
- additional information represented by macros (@)
- the list of possible redistributions (%active, %passive, etc.)

Conclusions and perspectives

Objectives

- Three major objectives
  1. **convert** Lexicon-Grammar tables to an NLP format,
  2. **plug** the resulting lexicon $LGLex_{\mathrm{Lefff}}$ with a parser
  3. **evaluate** the resulting parser
- parser used:
  - **FRMG** is a parser **TAG** of French [Thomasset & de La Clergerie 2005] from the compilation of a **meta-grammar**

Example of dependencies



Paul s'adresse à Max (*Paul talks to Max*)
→ entry *s'adresser V_33_8*

Long-term

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

- coding the **missing and uncoding entries** in the tables
- **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)

References

- [Constant & Tolone 2010] Matthieu Constant and Elsa Tolone. A generic tool to generate a lexicon for NLP from Lexicon- Grammar tables. Lingue d'Europa e del Mediterraneo, Grammatica comparata, vol.1, pp.79-93. Aracne. 2010.
- [Gross 1975] Maurice Gross. Méthodes en syntaxe : Régime des constructions complétives. Hermann. Paris, France. 1975.
- [Sagot 2010] Benoît Sagot. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. Proceedings of LREC'10, 8 pp. Valletta, Malta. 2010.
- [Thomasset & de La Clergerie 2005] François Thomasset and Éric de La Clergerie. Comment obtenir plus des méta-grammaires. Proceedings of TALN'05. Dourdan, France. 2005.
- [Tolone 2009] Elsa Tolone. Les tables du Lexique-Grammaire au format TAL. Proceedings of MajecSTIC'09. Avignon, France. 2009.
- [Tolone & Sagot 2011] Elsa Tolone and Benoît Sagot. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. LNAI. Springer Verlag. 2011. To appear.

Links

- phd tesis:
  `http://www-igm.univ-mlv.fr/~tolone/phd.pdf`
  $\rightarrow$ the version 3.3 will contain *LGLex* of tables in all formats for each category on `http://infolingu.univ-mlv.fr/` (Languages Ressources $>$ Lexicon-Grammar)
- post-doc in FaMAF of UNC, Córdoba :
  `http://www.cs.famaf.unc.edu.ar/~pln/`
- email: `elsa.tolone@univ-paris-est.fr`