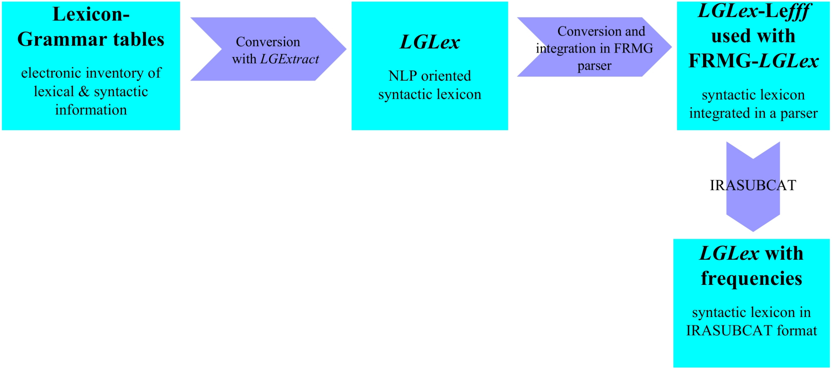
Frequencies of occurrence of entries and subcategorization frames in *LGLex* lexicon with IRASUBCAT

Elsa Tolone[1] & Romina Altamirano[1]

1. FaMAF, Universidad Nacional de Córdoba (Argentina)

LGC2013
Universidade do Algarve, Faculdade de Ciências Humanas e Sociais, Campus de Gambelas (Portugal)
September 10-14, 2013

Using IRASubcat with the converted lexicon and the relevant information extracted of the processed corpus we can complete the lexicon with the frequencies of occurrence for each verb and each syntactic function


Lexicon-Grammar tables
electronic inventory of lexical & syntactic information
Conversion with *LGExtract*
LGLex
NLP oriented syntactic lexicon
Conversion and integration in FRMG parser
LGLex*-Lefff used with FRMG-*LGLex
syntactic lexicon integrated in a parser
IRASUBCAT
***LGLex* with frequencies**
syntactic lexicon in IRASUBCAT format

1. Lexicon-Grammar tables for French

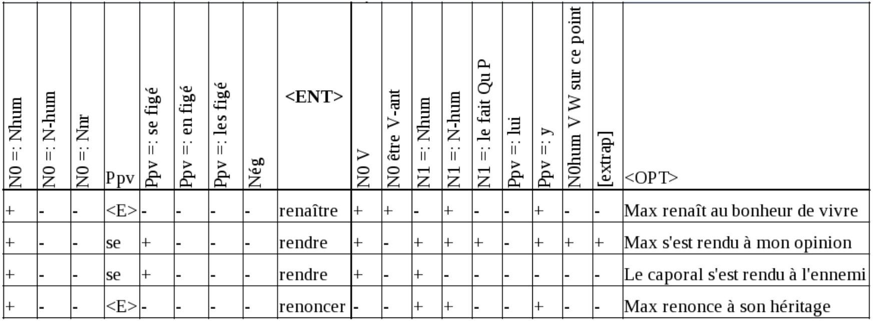
2. IRASUBCAT

3. Experiment with IRASUBCAT and the *LGLex* lexicon of French

4. Results

1. Lexicon-Grammar tables for French

Example: 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 in table of classes: N0 V à N1

[Gross 1975 ; 1994 ; LADL since 1970s ; LIGM since late 1990s]

LGLex

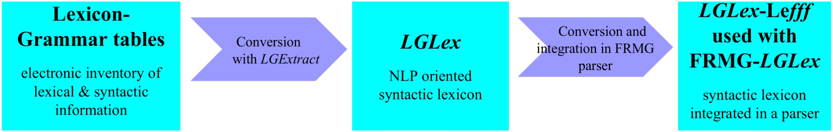
The improvement of the tables enables the extraction of a **syntactic lexicon** for each categories from Lexicon-Grammar tables [Constant & Tolone 2010]:

- named *LGLex* lexicon
- generated from the original Excel or CSV tables by the *LGExtract* tool
- exchange format with the same linguistic concepts of the tables
- text or XML format

The conversion towards the Alexina format enables the integration of them in a real-life **symbolic parser** [Tolone & Sagot 2011 ; Tolone *et al.* 2012]

- NLP tools used:
  - parser: FRMG [Thomasset & de La Clergerie 2005]
  - lexical formalism: Alexina, formalism used by the Le*fff* lexicon [Sagot 2010] used by FRMG
- named *LGLex*-Le*fff* lexicon → this allows a comparison between $\text{FRMG}_{\text{Le}\textit{fff}}$ and $\text{FRMG}_{\textit{LGLex}}$

Conversion of Lexicon-Grammar tables


Lexicon-Grammar tables
electronic inventory of lexical & syntactic information
Conversion with *LGExtract*
LGLex
NLP oriented syntactic lexicon
Conversion and integration in FRMG parser
LGLex*-*Lefff* used with FRMG-*LGLex
syntactic lexicon integrated in a parser

`http://infolingu.univ-mlv.fr/english` > Language Resources > Lexicon-Grammar > Download

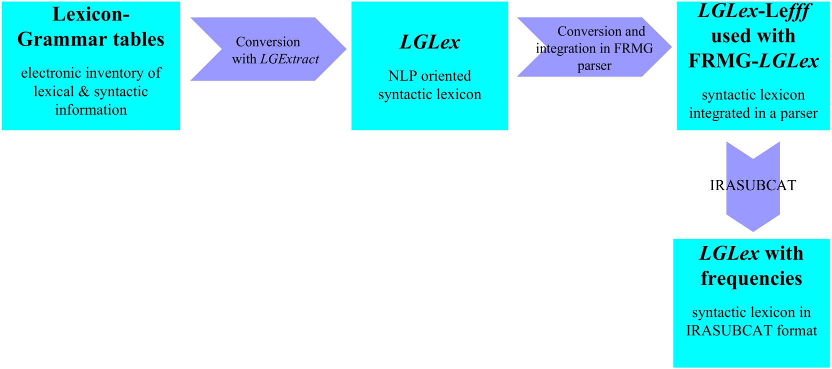
[Tolone 2012]

2. IRASUBCAT

IRASUBCAT

- a tool that acquires subcategorization information about the behaviour of any tag class (e.g., part of speech, syntactic function, etc.) or combination of them, from corpora
- takes as input a corpus in XML format
- the output is a lexicon, also in XML format, where each of the verbs under inspection is associated to a set of subcategorization patterns. The lexicon also provides information about frequencies of occurrence for verbs, patterns, and their co-occurrences in corpus
- allows to integrate the output lexicon with a preexisting one, merging information about verbs and patterns with information that had been previously extracted, possibly from a different corpus or even from a hand-built lexicon [Altamirano & Alonso Alemany 2010]

Adding frenquencies with IRASUBCAT



Lexicon-Grammar tables

electronic inventory of lexical & syntactic information

Conversion with *LGExtract*

LGLex

NLP oriented syntactic lexicon

Conversion and integration in FRMG parser

LGLex*-Lefff used with FRMG-*LGLex

syntactic lexicon integrated in a parser

IRASUBCAT

***LGLex* with frequencies**

syntactic lexicon in IRASUBCAT format

3. Experiment with IRASUBCAT and the *LGLex* lexicon of French

The experiment

We want to use the results of FRMG parser on a big corpus with IRASubcat in order to improve the *LGLex* lexicon of French, adding the frequencies of occurrence for each entry and each subcategorization frame. To do this, we must:

- choose a corpus with millons of words, also we just only need a small part of this corpus for the experiment
- parse the corpus with the FRMG parser, with and without the *LGLex* lexicon (i.e. only with the Le*fff* lexicon) – results with $\text{FRMG}_{\mathit{LGLex}}$ and with $\text{FRMG}_{\text{Le}\mathit{fff}}$
- convert both the processed corpus and the *LGLex* lexicon into XML format, required by IRASubcat;
- use IRASubcat in order to add the frequencies of occurrence extracted from the big corpus into the *LGLex* lexicon

The corpus

The processed corpus with FRMG$_{LGLex}$ to see how we use the FRMG parser with the *LGLex* lexicon) used for the experiment is the CPJ (Corpus Passage Jouet) with 100K sentences of AFP (Agence France-Presse), Europarl, Wikipedia and Wikisources, extracted from the corpus of the evaluation campaign (in 2009) for French parsers Passage [Hamon *et al.* 2008]

Conversion into XML format

We created 2 programs in Python:

- one to convert the verbal *LGLex* lexicon in the same format as IRASubcat output lexicon
- another to convert the processed corpus CPJ with the FRMG parser in a format directly readable by IRASubcat

Conversion of the verbal *LGLex* lexicon

- The input is the verbal *LGLex* lexicon, or more precisely, the *extensional lexicon* of *LGLex*-Lefff lexicon, which contains each inflected form of the lemma and every possible redistribution
- In the output lexicon converted into XML format as IRASubcat output lexicon (named *lglex-lefff-IRASubcat.xml*), each lemma is associated to a set of subcategorization patterns. For example:
  <pattern id="['Suj:cln|sn', 'Obj:sn']"></pattern>
  <pattern id="['Suj:(cln|sn)', 'Obl:de-sinf']"></pattern>

Conversion of the verbal *LGLex* lexicon: An example

We have in total 14,068 distinct lemmas.

Here is a complete example of *lglex-lefff-IRASubcat.xml*:

```
<dictionary>
  <entry verb="achever__ _V_1_1" count_oc_verb="0">
    <tag name="fs" different_patterns="6">
      <pattern id="['obj', 'suj']" count_w_verb="0" total_count="0"
        rejected_patterns_freq_test="NO"></pattern>
      <pattern id="['obl', 'suj']" count_w_verb="0" total_count="0"
        rejected_patterns_freq_test="NO"></pattern>
      <pattern id="['obl2', 'suj']" count_w_verb=="0" total_count="0"
        rejected_patterns_freq_test="NO"></pattern>
      <pattern id="['obl', 'obl2']" count_w_verb=="0" total_count="0"
        rejected_patterns_freq_test="NO"></pattern>
    </tag>
  </entry>
</dictionary>
```

Conversion of the processed corpus with the FRMG parser(1)

- The input is the processed corpus CPJ with the FRMG parser, more precisely, with FRMG$_{LGLex}$, i.e. the FRMG parser with the *LGLex*-Le*fff* lexicon. In the processed corpus CPJ, we represent in XMLDep format a graph of dependencies with nodes (lemmas), grouped in clusters (forms), with arcs describing the syntactic dependencies between nodes. So, we want to extract only the useful information in a format directly readable by IRASubcat

Conversion of the processed corpus with the FRMG parser(2)

- In the output in XML format (named *CPJ-IRASubcat.xml*), for each sentence of the corpus (for example, <**sentence ID="12" corpus="frwikipedia_012" s="12"**>), we extracted the verbs (**cat="v"**) with their identifiers (for example, **lemmaid="achever___V_1_1**"). For each verb, we extracted the syntactic functions and we indicated the number of arguments (**nb_fs="2"**) and then, each syntactic function (**fs**) one by one (for example, **fs="suj"** for subject, and **fs="obl2"** for oblique)

Conversion of the processed corpus with the FRMG parser: An example

Here is a complete example of *CPJ-IRASubcat.xml*:

```
<sentence ID="12" corpus="frwikipedia_012" s="12">
   <word lexica="achevée" lemma="achever" lemmaid="achever__V_1_1"
     cat="v" nb_fs="2">achevée</word>
   <word fs="suj"></word>
   <word fs="obl2"></word>
</sentence>
```

Using IRASubcat with *LGLex*

We changed the information in the configuration file to execute IRASubcat with our lexicon *lglex-lefff-IRASubcat.xml* and our corpus *CPJ-IRASubcat.xml* (in UTF-8):

```
VERB LIST = NO
EXISTING DICTIONARY = lglex-lefff-IRASubcat.xml
LENGTH OF VERBAL CONTEXT = 3
COMPLETE WITH EMPTY WORD = NO
KEEP ORDER = NO
TARGET TAGS = fs
USE LEXICAL FORM OF WORDS = NO
INTRODUCE VERBAL MARK = NO
COLLAPSE PATTERNS = NO
MAX ITERATION TO COLLAPSE PATTERNS = FALSE
MIN FREQUENCY OF VERBS = 0
MIN REL FERQUENCY OF PATTERNS = 0
USE LIKEHOOD RATIO TEST = NO
```

4. Results

The execution

The execution create:

- the file *OutputDictionaryOrd.xml* with the lexicon
- the file *info_file* with the statistics of execution
- the file *IdsSentencesOrigenDictionary.xml* with the ID's of sentences that give origin of the patterns in *OutputDictionaryOrd.xml*

The result lexicon (1)

Here is the previous example of *lglex-lefff-IRASubcat.xml* as it appears in *OutputDictionaryOrd.xml*:

```
<dictionary>
  <entry verb=”achever__V_1_1” count_oc_verb=”1”>
    <tag name=”fs” different_patterns=”4”>
      <pattern id=”[’obj’, ’suj’]” count_w_verb=”0” total_count=”1001”
        rejected_patterns_freq_test=”NO”></pattern>
      <pattern id=”[’obl’, ’suj’]” count_w_verb=”0” total_count=”214”
        rejected_patterns_freq_test=”NO”></pattern>
      <pattern id=”[’obl2’, ’suj’]” count_w_verb=”1” total_count=”325”
        rejected_patterns_freq_test=”NO”></pattern>
      <pattern id=”[’obl’, ’obl2’]” count_w_verb=”0” total_count=”0”
        rejected_patterns_freq_test=”NO”></pattern>
    </tag>
  </entry>
</dictionary>
```

The result lexicon (2)

- We can see that the number of occurrences of the verb **achever__V_1_1** in the corpus is 1 and the pattern is **['obl2', 'suj']**. For this pattern, we have in total 325 occurences in the corpus for all verbs
- We can see in the example of *IdsSentencesOrigenDictionary.xml* (see below) that the occurence of **verb="achever__V_1_1"** with the pattern **['obl2', 'suj']** is in the sentence **['12']**

The result lexicon (3)

```
<ids_from>
  <entry verb=”achever__V_1_1” total_count=”1”>
    <tag name=”fs”>
      <pattern id=”[’obj’, ’suj’]”>
        <s_list>[]</s_list>
      </pattern>
      <pattern id=”[’obl’, ’suj’]”>
        <s_list>[]</s_list>
      </pattern>
      <pattern id=”[’obl2’, ’suj’]”>
        <s_list>[’12’]</s_list>
      </pattern>
      <pattern id=”[’obl’, ’obl2’]”>
        <s_list>[]</s_list>
      </pattern>
    </tag>
  </entry>
</ids_from>
```

The frequencies: Number of occurrences of patterns

The frequencies indicated in *OutputDictionaryOrd.xml* allow us to know the total number of occurences of each pattern in the corpus. We don't indicate the patterns which never appear

pattern	*total_count*
['obj', 'suj']	1001
['obl2', 'suj']	325
['obl', 'suj']	214
['att', 'suj']	142
['loc', 'suj']	92
['objà', 'suj']	91
['suj']	62
['objde', 'suj']	55
['obj']	26
['dloc', 'suj']	11
others	0

The frequencies: Number of occurrences of verbs

The frequencies indicated in *IdsSentencesOrigenDictionary.xml* allow us to calculate the number of verbs associated with each total number of occurences of this verbs. We indicate the verb when there is only one verb

verb or nb of verbs	*total_count*	nb of verbs	*total_count*
être_____2	63	4	9
pouvoir___V_1_88	60	3	8
devoir___V_1_38	37	8	7
faire_____2	22	12	6
dire___V_9_130	19	14	5
vouloir___V_15_82	17	30	4
2	16	63	3
avoir___V_37E_10	13	192	2
2	12	740	1
3	10	13 043	0

Conclusions and perspectives

- The processed corpus is the results of the FRMG parser with *LGLex* lexicon, so it could find wrong sense
- The next step is to consider the information on realizations, that we must extract from processed corpus, but it is not a straightforward task
- Then we have to use the FRMG parser with Le*fff* lexicon only, without the *LGLex* lexicon influences the results
- We could also use IRASubcat with another parser which is statistical, such as MaltParser, MSTParser, or Berkeley Parser Candito *et al.* 2010]
- And we could do a comparison using the original lexicon and the enlarged lexicon with that different parsers to verify that the accuracy is better using more information

References (1)

- [Altamirano & Alonso Alemany 2010] Altamirano R. & Alonso Alemany L. IRASUBCAT, a highly parametrizable, language independent tool for the acquisition of verbal subcategorization information from corpus. Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pp. 84-91, Los Angeles, California, 2010.
- Candito *et al.* 2010] Candito M., Nivre J., Denis P. & Henestroza Anguiano E. Benchmarking of statistical dependency parsers for french. Proceedings of COLING'2010 (poster session), Beijing, China, 2010.
- [Constant & Tolone 2010] Constant M. & Tolone E. 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.

References (2)

- [Gross 1975] Gross M. Méthodes en syntaxe : Régime des constructions complétives. Hermann. Paris, France. 1975.
- [Gross 1994] Gross M. Constructing Lexicon-Grammars. Oxford University Press, Oxford, England, 1994.
- [Hamon *et al.* 2008] Hamon O., Mostefa D., Ayache C., Paroubek P., Vilnat A. & de La Clergerie E. Passage: from French parser evaluation to large sized treebank. Proceedings of LREC'08, Marrakech, Morocco, 2008.
- [Sagot 2010] Sagot B. The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for French. Proceedings of LREC'10, 8 pp. Valletta, Malta. 2010.

References (3)

- [Thomasset & de La Clergerie 2005] Thomasset F. & de La Clergerie E. Comment obtenir plus des méta-grammaires.Proceedings of TALN'05. Dourdan, France. 2005.
- [Tolone & Sagot 2011] Tolone E. & Sagot B. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. LNAI. Springer Verlag. 2011.
- [Tolone 2012] Tolone E. Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français, Editions Universitaires Européennes, Saarbrücken, Germany, 352 pp. 2012.
- [Tolone *et al.* 2012] Tolone E., Sagot B. & Eric de La Clergerie. Evaluating and improving syntactic lexica by plugging them within a parser. Proceedings of LREC'12, 8pp. Istambul, Turquia. 2012.