Conversión de las tablas del Léxico-Gramática del francés en el léxico *LGLex*

Elsa Tolone[1]

1. LIGM, Université Paris-Est (Francia) & FaMAF, Universidad Nacional de Córdoba (Argentina)

Seminar of PLN group – InCo, Facultad de Ingeniería, Universidad de la República, Uruguay
October 27, 2011

Context in 1970's

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

Objectives

- Three major objectives
  1. **convert** Lexicon-Grammar tables to an NLP format,
  2. **plug** the resulting lexicon, named $LGLex_{\mathrm{Lefff}}$, with a parser
  3. **evaluate** the resulting parser
- NLP tools used:
  - parser: FRMG [Thomasset & de La Clergerie 2005]
  - lexical formalism: Alexina, formalism used by the Lefff lexicon [Sagot 2010] used by FRMG

$\rightarrow$ this allows a comparison between $\mathrm{FRMG}_{\mathrm{Lefff}}$ and FRMG_{LGLex}

1. Lexicon-Grammar tables for French

2. *LGLex* lexicon

3. Conversion of *LGLex* into the Alexina format

1. Lexicon-Grammar tables for French

Lexicon-Grammar tables

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

Methodology:

- Study the syntax of basic sentences (or **subcategorization frames**)
  e.g.: N0 V N1
- Study of French verbs, adverbs, predicative nouns and adjectives and frozen expressions
  → they **share some features** = **classes**
- The different meanings are distinguished
  e.g.: *se rendre* (*to surrender* / *to accept*
  - *rendirse* : *capitular* / *aceptar*)

Principle

- Each class is described in a **table**:
  - one row for each (lemma-level) entry
  - one column for each feature relevant to the class
  - in each cell, $+$ (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 include:
  - a **basic defining** feature (a subcategorization frame)
  - often additional features (distributional, morphological, transformational, semantic, etc.)
    e.g.: N0 =: Nhum $\rightarrow$ human name

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

Inventory

- Inventory [Tolone 2009]:
  - 67 classes of **simple verbs**
    - 13 872 entries for 5 738 distinct lemmas
  - 81 classes of **simple and compound predicative nouns** (nouns with argument(s)that are studied with their light verb)
    - 14 271 entries for 10 112 distinct lemmas
      e.g.: *Luc monte une attaque contre le fort*
      (*Luc is lauching an attack against the fort*)
  - 69 classes of **frozen expressions**, mostly verbal and adjectival
    - 39 628 entries for 38 658 distinct lemmas
      e.g.: *Tu n'arrives pas à la cheville de Marie* (*You don't hold a candle to Mary - no llegarle a los talones / a la suela del zapato*, literaly *You don't arrive at the ankle of Mary - no llegarle al tobillo*)
  - 32 classes of **simple** (adverbs in *-ment*) **and frozen adverbs**
    - 10 488 entries for 9 326 distinct lemmas
      e.g.: *[changer] du jour au lendemain* (*[to change] overnight - [cambiar] de un día para otro*)

Problems

- **Different names** for the same feature
  → Harmonization of the column headings
  ex : [extrap] and il V N0 W
- Features **not defined clearly**
  → Documentation of features
- **Implicit defining features** (literature)
  → Constant + or − for the whole table
- **Outdated** entries in the tables
  → Adding missing entries
- **Uncoded** entries (∼)
  → Coding entries

3. *LGLex* lexicon

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 version 3.4 contained *LGLex* for each category at `http://infolingu.univ-mlv.fr/`

Format of *LGLex* lexicon

- **ID=category_numTable_numEntry**
- **lexical-info**=[...] → lemma and lexical information (auxiliaries, support verbs, determiners, prepositions)
- **args**=(...) → arguments and their nature with other information (semantic features, mood of the complementizer phrase, argument controled by the infinitive, prepositions)
- **all-constructions**=[...] → list of accepted constructions
- **example**=[...] → an illustrative example of the entry

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*)

3. Conversion of *LGLex* into the Alexina format

The FRMG parser

A **TAG** parser of French [Thomasset & de La Clergerie 2005]
FRMG fits into a processing chain:

- upstream
  - **SXPipe**: segmentation, token, corrections, named entities
  - **Lefff**: morphological and syntactic lexicon for French
    $\rightarrow$ connection lexicon/grammar: anchoring with **hypertag**
- downstream, with a module of **disambiguation** (with heuristics)

The Lefff

- 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)
- 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

- **Syntactic functions** (cf. Dicovalence): Suj, Obj, Objà, Objde, Loc, Dloc, Att, Obl/Obl2
- **Realizations**: direct (sn, sa, sinf, scompl, qcompl); clitic (cln, cla, cld, y, en); prepositional (prep+direct, e.g., par-sn, à-sinf, de-scompl)
- **Redistributions**: %active, %passive, etc.

Conversion of *LGLex*

- The conversion of Lexicon-Grammar tables into the Alexina framework is **not straightforward** [Tolone & Sagot 2011]
  - It requires a **formal definition** or a **dynamic interpretation** of all feature names
- We won't enter into the details of this conversion process
  - The version 3.4 contained *LGLex*-Lefff for verbs and predicatives nouns at `http://infolingu.univ-mlv.fr/`

LGLex: the previous example

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$_{\text{Lefff}}$: 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$_{\text{Lefff}}$: 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.)

The resulting lexicon: *LGLex*$_{\text{Lefff}}$

- **verbes** : *LGLex*$_{\text{Lefff}}$ contains 22 128 entries for 5 739 unique verb lemmas (3,85 entries per lemma)
  - to be compared with the last published version of the Lefff: 7 072 verb entries for 6 818 unique verb lemmas (1,04 entries per lemma)
- **predicatives nouns** : *LGLex*$_{\text{Lefff}}$ contains 30 443 entries for 10 069 distinct lemmas (3,02 entries per lemma)
  - The Lefff contains only 218 entries of predicative nouns (1 entry per lemma)

Conclusions and perspectives

Version 3.4 of tables of Lexicon-Grammar

All tables available under LGPL-LR license at
`http://infolingu.univ-mlv.fr/`

- Simple verbs :
  - 67 tables + the table of classes (552 features)
  - an index of all entries
  - documentation of features
  - defining formulas of each table
  - classification tree
- Simple and compound predicative nouns :
  - 81 tables + the table of classes (516 features)
- Frozen expressions :
  - 69 tables + the table of classes (276 features)
- Simple and frozen adverbs:
  - 32 tables + the table of classes (159 features)

The pursuit of the method

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

- continue to **improve the tables**
- coding the **missing and uncoding entries** in the tables
- coding the **table of classes** for each category
- **improve/correct the conversion process**
- apply this technique to Lexicon-Grammar tables for **other categories**
- generalize this technique to **other languages** for which large-coverage Lexicon-Grammar tables are available (e.g., Greek)

Improve FRMG$_{LGLex}$ and *LGLex*

- Improve FRMG$_{LGLex}$:
  - **coupling both parser variants** could prove useful, since full parses have a higher f-measure than partial parses
  - **detecting errors**: use automatic techniques, e.g., error mining of parsing results [Sagot & de La Clergerie 2008])
- Improve the resulting lexicon :
  - deducing the **weights** with an annotated corpus in order to distinguish unusual uses
  - **merging** with other lexical ressources
  - using of the lexicon in other systems of NLP

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.
- [Sagot et de La Clergerie 2008] Benoît Sagot and Éric de La Clergerie. Fouille d'erreurs sur les sorties d'analyseurs syntaxiques. Traitement Automatique des Langues (T.A.L.), vol.49, num.1, pp. 41-60. Hermès. Paris, France. 2008. 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.

References (2)

- [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 et al. 2010] Elsa Tolone, Stavroula Voyatzi and Christian Leclère. Constructions définitoires des tables du Lexique-Grammaire. Actes de LGC'10, pages 321-331. Belgrade, Serbie. 2010.
- [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 Thesis : `http://www-igm.univ-mlv.fr/~tolone/phd.pdf`
- Ressources : `http://infolingu.univ-mlv.fr/`
  > Languages Ressources > Lexicon-Grammar > Download
- Email: `elsa.tolone@univ-paris-est.fr`

Versions

Rich and large-coverage lexicon available on
`http://infolingu.univ-mlv.fr/`

- version 1 = origin format
- version 2 = first tables online (only 60%)
- version 3 = since my tesis, in the different formats of lexicon
  - tables → lisibility for the linguists
  - *LGLex* → exchange format
  - $LGLex_{\mathrm{Lefff}}$ → format to integrate in a parser

Features documentation

Feature	Description of the feature
N0 =: homme	actividad donde el sujeto practicante es exclusivamente masculino
Det pluriel obl	el determinante es obligatoriamente en plural
A (Prép+E) B = B	indica para los nombres compuestos construidos con dos componentes (A,B) que el primero componente (A) puede borrarse
élément dérivé	indica para los nombres compuestos sobre qué elemento se aplica la derivación
suff-	sufijo quitado al nombre predicativo para formar el Nagent (Nagento) derivado
verbe associé	indica que al nombre predicativo está asociado un verbo
N0 jouer à Det N	indica que en paralelo a la frase *faire de "activité"* (*hacer de "actividad"*) existe una frase *jouer à "activité"* (*jugar a "actividad"*)

The FRMG parser

FRMG is a **TAG** parser of French [Thomasset & de La Clergerie 2005]

- from the compilation of a **meta-grammar**
- very **compact** thanks to the factorization of trees (1 986 300 defactorized trees $\rightarrow \sim$ 75 630 trees)
- exploiting the functionalities of **DyALog** $\rightarrow$ logic programming environment

FRMG fits into a processing chain:

- upstream
  - **SXPipe**: segmentation, token, corrections, named entities
  - **Lefff**: morphological and syntactic lexicon for French $\rightarrow$ connection lexicon/grammar: anchoring with **hypertag**
- downstream, with a module of **disambiguation** (with heuristics)

The conversion process

- The conversion of Lexicon-Grammar tables into the Alexina framework is **not straightforward** [Tolone & Sagot 2011]
  - It requires a **formal definition** or a **dynamic interpretation** of all feature names
    Example :
    - N0 V N1 → basic defining feature
    - N0 V → erasement of N1
    - N0 V N1 à N2 → elongation of the basic defining feature
    - N0 V Qu P → realization of N1
    - [passif par] → passive redistribution
  - Additional important information must be gathered heuristically or from other lexical resources
    - the name of each syntactic function, attribution phenomena, morphological information, etc.

5. Integration in the FRMG parser

Integration in the FRMG parser

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

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

6. Evaluation and discussion

Protocol used

- We evaluated $\text{FRMG}_{\text{Lefff}}$ and FRMG_{LGLex} by parsing the manually annotated part of the 1st Passage parsers' evaluation campaign of 2007 [Hamon *et al.* 2008]
  - 4 306 sentences of EASy annotated corpus + 400 new sentences : 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** ($\sim$ 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
  $\rightarrow$ this conversion process certainly introduces errors
- the majority of predicative nouns can not evaluated because FRMG does not consider those with determiners
- Passage not allow to evaluate all the information contained in tables (e.g. semantic features)
- the Lefff was developed in parallel with the EASy and Passage campaigns (unlike Lexicon-Grammar tables)
- *LGLex*$_{\text{Lefff}}$ does not contain all necessary verb entries; we added other ones
  $\rightarrow$ other verb entries may be still missing because all verb entries are not encoded

Results

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

	Chunks		Relations	
Sub-corpus	$\text{FRMG}_{\text{Lefff}}$	FRMG_{LGLex}	$\text{FRMG}_{\text{Lefff}}$	FRMG_{LGLex}
`general_lemonde`	**88.22%**	84.60%	**62.73%**	59.01%
`litteraire_2`	**88.91%**	88.46%	**65.28%**	62.43%
`mail_9`	**82.60%**	81.90%	**58.55%**	56.00%
`medical_3`	85.04%	**85.89%**	64.79%	**65.26%**
`oral_delic_4`	78.80%	**81.79%**	**51.67%**	51.14%
`questions_amaryllis`	**91.30%**	90.73%	**66.56%**	64.77%
total	**87.05%**	*85.53%*	**63.10%**	*60.25%*

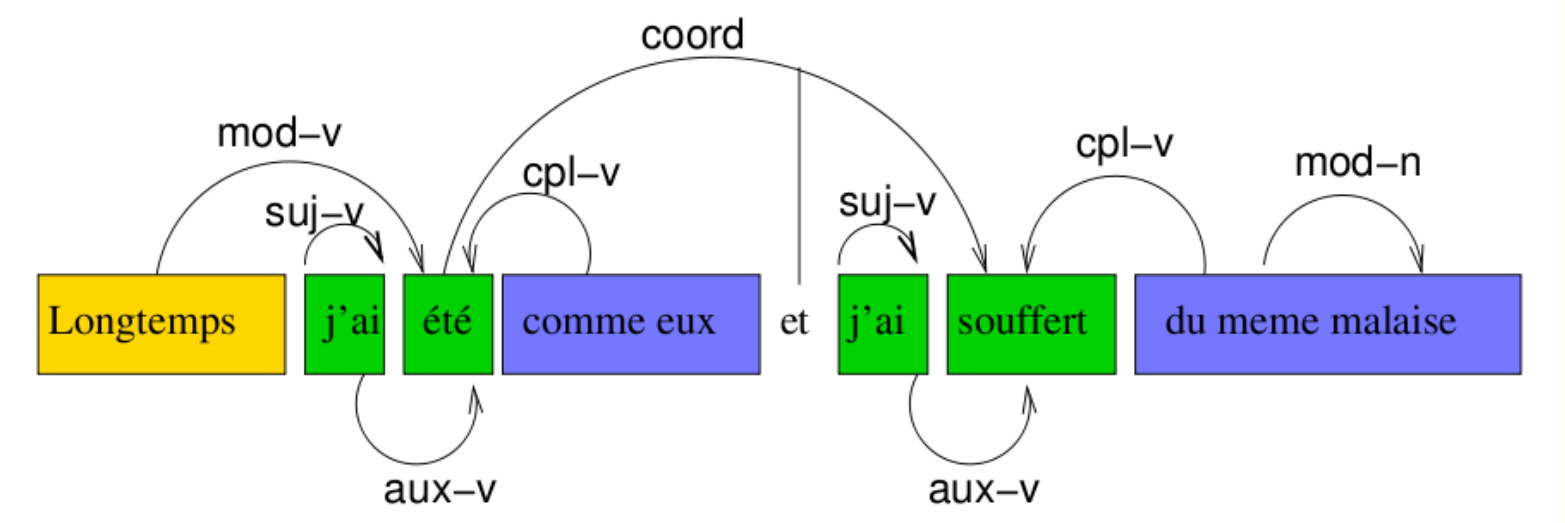
Parsing times higher with FRMG_{LGLex} than with $\text{FRMG}_{\text{Lefff}}$: the median parsing time per sentence is 0,62s vs. 0,26s

- this comes from the higher average number of entries per verb lemma (approx. 3) in *LGLex* than in the Lefff
  $\rightarrow$ more ambiguity

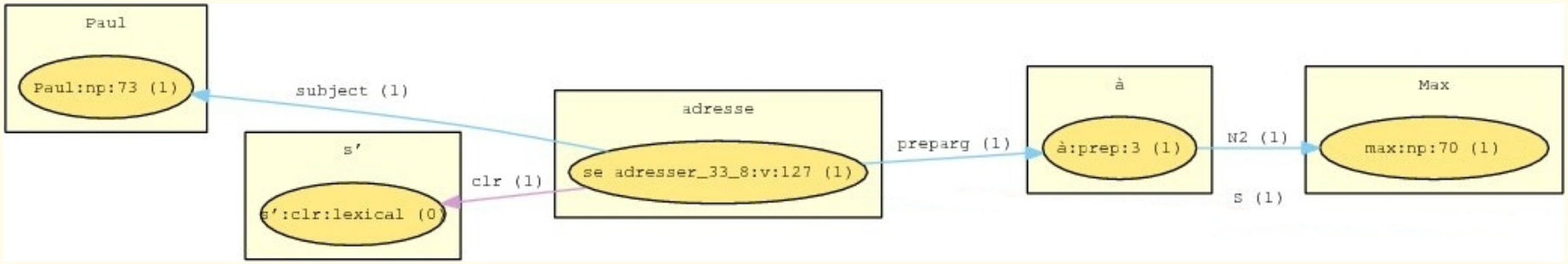
- FRMG_{LGLex} gives better results than $\text{FRMG}_{\text{Lefff}}$ for some relations
  - "standard" relations MOD-A (adjective modifier) and MOD-R (adverb modifier)
  - "tough" relations MOD-P (preposition modifier) and APP (apposition)
- the ATB-SO relation (subject or object attribute) is the relation 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, but it can be due to errors of reference

- the higher **lexical ambiguity** in FRMG_{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_{LGLex} considers *de paix* as an argument of *estimer* (*estimer qqch de qqn*)
  - $\text{FRMG}_{\text{Lefff}}$ makes no error since in the Lefff, *estimer* has no Objde

Example: annotation of relations



Example of dependencies in FRMG



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

References (3)

- [Hamon et al. 2008] Hamon O., Mostefa D., Ayache C., Paroubek P., Vilnat A. and La Clergerie E. Passage: from French Parser Evaluation to Large Sized Treebank. Proceedings of LREC'08. Marrakech, Maroc. 2008.
- [Paroubek et al. 2006] Patrick Paroubek, Isabelle Robba, Anne Vilnat and Christelle Ayache. Data, Annotations and Measures in EASy: the Evaluation Campaign for Parsers of French. Proceedings of LREC'06. Genoa, Italy. 2006.