Evaluating and improving syntactic lexica by plugging them within a parser

Elsa Tolone, Benoît Sagot, Éric Villemonte de La Clergerie

Evaluating and improving syntactic lexica by plugging them within a parser

Elsa Tolone[1], Benoît Sagot[2], Éric Villemonte de La Clergerie[2]

1. FaMAF, Universidad Nacional de Córdoba, Medina Allende s/n, Ciudad Universitaria, Córdoba, Argentina
2. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France
elsa.tolone@univ-paris-est.fr,{benoit.sagot,eric.de_la_clergerie}@inria.fr

Abstract

We present some evaluation results for four French syntactic lexica, obtained through their conversion to the Alexina format used by the Le*fff* lexicon (Sagot, 2010), and their integration within the large-coverage TAG-based FRMG parser (de La Clergerie, 2005). The evaluations are run on two test corpora, annotated with two distinct annotation formats, namely EASy/Passage chunks and relations and CoNLL dependencies. The information provided by the evaluation results provide valuable feedback about the four lexica. Moreover, when coupled with error mining techniques, they allow us to identify how these lexica might be improved.

Keywords: syntactic lexica, parsing, error mining

1. Introduction

The development of a large-scale symbolic parsing system is generally achieved by co-developing its various components in a consistent way, in order to ensure good integration and consistency. This is the case, for example, concerning the development of the French syntactic lexicon Le*fff* (Sagot, 2010) together with that of the FRMG grammar and parser for French (de La Clergerie, 2005).

However, several syntactic lexica have been developed for French, sometimes for decades. An example thereof are Lexicon-Grammar tables (Gross, 1975; Boons et al., 1976b), although they were built with no or poor integration in large-scale NLP systems. Such a resource contains rich and valuable information, but its usefulness for real NLP applications has never been thouroughly assessed. This is also the case for the valency lexicon DICOVALENCE (van den Eynde and Mertens, 2006), although to a lesser extent.

Therefore, after converting these resources in the same lexical model as the Le*fff*, we performed a preliminary integration of these resources within FRMG. We also evaluated NewLe*fff*, a new experimental version of the Le*fff* that benefits, among other things, from its merging with DICOVALENCE (Sagot and Danlos, 2012). Following previous preliminary results (Tolone et al., 2011), we were able to evaluate these lexica on two reference corpora based on two different annotation schemes, namely the EasyDev corpus (Paroubek et al., 2009) and on the CONLL dependency version of the French TreeBank (Candito et al., 2010). As a side effect, these experiments also show that it is possible to switch lexica in a lexicalized parser like FRMG, at the cost of a relatively small decrease in performances.

2. Lexica

2.1. The Le*fff*

Our reference and baseline lexicon is Le*fff* (Sagot, 2010), a large coverage morphosyntactic and syntactic lexicon for French.[1] As mentioned before, Le*fff* was specifically developed for NLP tasks, and in particular to be used in conjunction with the FRMG parser.

The Le*fff* relies on the Alexina framework for the acquisition and modeling of morphological and syntactic lexica. To represent lexical information, an Alexina lexicon relies on a two-level architecture:

- the *intensional* lexicon associates (among others) an inflection table and a canonical sub-categorization frame with each entry and lists all possible redistributions from this frame;
- the *compilation* of the intensional lexicon into an *extensional lexicon* builds different entries for each inflected form of the lemma and every possible redistribution.

The current version of the Le*fff* (version 3.1)[2] contains only one entry for the lemma *vérifier* 'verify', 'validate'. Here is a simplified version of this entry:

vérifier$_1$ *Lemma;v;<Suj:cln|sn,*
Obj:(cla|qcompl|scompl|sinf|sn)>;
@CtrlSujObj, @CompSubj cat=v;
%ppp_employé_comme_adj,%actif,%passif,
%se_moyen_impersonnel,%passif_impersonnel

It describes a transitive verb entry whose arguments have the *syntactic functions* *Suj* and *Obj* listed between angle brackets.[3] The subject might be realized as a nominative clitic (*cln*) or a noun phrase (*sn*), whereas the direct object can be realized as an accusative clitic (*cla*), a noun phrase (*sn*), an infinitive (*sinf*, with a control phenomenon expressed by *@CtrlSujObj*), a finite clause (*scompl*, in the subjunctive mood because of *@CompSubj*) or an indirect

[1] Freely available at http://gforge.inria.fr/projects/alexina/

[2] The Le*fff* 3.1 package includes `v_new` files that are not considered as being part of the Le*fff* yet. In fact, replacing `v` and `v-phd` files by `v_new` files leads to what is called NewLe*fff* in this paper. See below for more details.

[3] The different syntactic functions used in the Le*fff* are: *Suj* (subject), *Obj* (direct object), *Objà* (indirect object canonically introduced by preposition "à"), *Objde* (indirect object canonically introduced by preposition "de"), *Loc* (locative), *Dloc* (delocative), *Att* (attribute), *Obl* or *Obl2* (other oblique arguments).

interrogative clause (*qcompl*). Finally, this verb entry allows for the functional redistributions *past participle used as an adjective*, *active* (the default distribution), *impersonal middle-voice "se" construction*, *impersonal passive*, and *passive*.
The Le*fff* 3.1 contains 7,108 verbal entries corresponding to 6,827 distinct lemmas, and 112,118 entries covering all other categories. Detailed figures are given in Table 1.[4]

2.2. Other syntactic lexica

Besides Le*fff*, we have considered three other lexica, whose verbal entries are combined with the non-verbal Le*fff* entries:

- *LGLex*:[5] this lexicon results from a two-step conversion of the Lexicon-Grammar tables (Gross, 1975; Boons et al., 1976b), a rich syntactic lexical resource developed over several decades although not with an NLP orientation. A first conversion was made to get a fully electronic version of this lexicon into the *LGLex* format (Constant and Tolone, 2010), opening the way to a second conversion to the Alexina format (Tolone and Sagot, 2011). The result is a wide-coverage lexicon with, often, many entries for each verbal lemma, associated to several meanings and valency frames. *LGLex* contains 13,867 verbal entries corresponding to 5,738 distinct lemmas, as well 12,696 entries for predicative nouns corresponding to 8,531 distinct nominal lemmas;
- DICOVALENCE:[6] this lexicon (van den Eynde and Mertens, 2006) follows the Pronominal Approach (Blanche-Benveniste et al., 1984) for characterizing verb valency frames and for defining fine-grained entries (several entries per lemma). This medium-coverage resource contains 8,313 verbal entries corresponding to 3,738 distinct lemmas;
- NewLe*fff*:[7] this experimental new version of Le*fff* targets more semantically-oriented finer-grained entries, while still preserving Le*fff*'s wide coverage. This lexicon is the result of two extension steps: (1) the automatic extraction, interpretation, conversion and integration or merging of denominal and deadjectival verbal entries in *-iser* and *-ifier* from the LVF lexicon (Sagot and Fort, 2009); (2) the automatic merging of Le*fff* entries and DICOVALENCE entries through the comparison of their valency frames (Sagot and Danlos, 2012), then completed by a phase of manual validation on the 100 most frequent lemmas and on all *dubious* lemma[8] (those lemma who got more entries than originally in both input lexica). NewLe*fff* contains 12,613 verbal entries corresponding to 7,933 distinct lemmas.

Table 2 provides some figures about the coverage and granularity of each lexicon for verbal entries, showing that our four lexica actually cover an diverse spectrum of configurations (medium to large coverage, small to large granularity). For the experiments, all lexica use the non-verbal entries of Le*fff* in addition of their own verbal entries. In consequence, the differences between the lexica arise from the verbal entries.

Lexica	#Entries	#Lemmas	Ratio
Le*fff*	7,108	6,827	1.04
LGLex	13,867	5,738	2.41
DICOVALENCE	8,313	3,738	2.22
NewLe*fff*	12,613	7,933	1.58

Table 2: All lexica at a glance (verbal entries)

3. FRMG

FRMG[9] (de La Clergerie, 2005) is a large-coverage symbolic grammar and parser for French. In fact, the acronym FRMG denotes resources that cover several representation levels. The most abstract level corresponds to a linguistically motivated modular and hierarchical meta-grammar. This meta-grammar is used to generate a compact (feature-based) Tree Adjoining Grammar (TAG) (Joshi et al., 1975) containing around 300 factorized elementary trees, including only 35 trees anchored by verbs. Despite its compactness, the grammar exhibits a wide coverage, thanks to factorization operators, such as disjunction and guards, used in the trees to allow many possible tree traversals.
The grammar is compiled into an efficient chart-based parser, also named FRMG, which is able to return both full parses (whenever possible) or sequences of partial parses (otherwise) as shared dependency forests. The forests may then be disambiguated using heuristic-based rules to get the best dependency trees. Finally, these trees may be converted to various output formats, including the EASy format and the CONLL format.
FRMG benefits from the extended domain of locality provided by TAG trees, with, for instance, the possibility to capture all the components of a verb valency frame through the nodes of a single elementary tree. However, it also implies that a TAG grammar like FRMG works best when coupled with a lexicon that provides such rich lexical information. There is also the need to propagate this information from the words to the trees. Concretely, each tree of FRMG is associated with an *hypertag* (Kinyon, 2000), a feature-structure resuming (in the case of a verbal tree) the various frames and argument realizations covered by the tree. Similarly, each verbal entry (but this is also true for other categories) has an hypertag derived from its lexical information. Anchoring a verbal tree by a verbal entry involves the unification of both hypertags. Figure 1 shows the

[4]The category "other" includes all kinds of conjunctions, determiners, interjections, punctuation marks, pronouns, prefixes and suffixes, as well as special entries for named entities and unknown words.

[5]Freely available at http://infolingu.univ-mlv.fr/english/ > Language Resources > Lexicon-Grammar > Download

[6]Freely available at http://bach.arts.kuleuven.be/dicovalence/

[7]Freely available at http://gforge.inria.fr/projects/alexina/. For obtaining NewLe*fff*, `v_new` files in the Le*fff* 3.1 package must be compiled and used instead of `v` and `v-phd` files.

[8]505 verbal lemmas, corresponding to 986 entries.

[9]Freely available at http://mgkit.gforge.inria.fr/

Category	#Intensional Entries	#Unique Lemmas	#Extensional Entries
verbs	7,108	6,827	363,120
verbal idioms	1,869	1,851	3,296
nouns	41,816	41,592	86,675
adjectives	10,556	10,517	34,359
adverbs	4,111	3,676	4,155
prepositions	260	259	728
proper nouns	52,499	52,202	52,571
other	1,007	854	1,589
total	119,226	117,778	546,493

Table 1: Quantitative data about the L*efff*

hypertag associated with one of the entries for *promettre* 'promise' corresponding to the frame "`(arg0)` promises `(arg1) to-(arg2)`" with a control of (optional) `arg1` by the subject `arg0`, and nominal or clausal realization for the (optional) object argument `arg1`.

$$\begin{bmatrix} \texttt{arg0} & \begin{bmatrix} \texttt{fun} & \text{suj} \\ \texttt{kind} & \text{subj} \mid \text{-} \\ \texttt{pcas} & \text{-} \end{bmatrix} \\ \texttt{arg1} & \begin{bmatrix} \texttt{fun} & \text{obj} \\ \texttt{kind} & \text{obj} \mid \text{scomp} \mid \text{-} \\ \texttt{pcas} & \text{-} \end{bmatrix} \\ \texttt{arg2} & \begin{bmatrix} \texttt{fun} & \text{objà} \\ \texttt{kind} & \text{prepobj} \mid \text{-} \\ \texttt{pcas} & \text{à} \mid \text{-} \end{bmatrix} \\ \texttt{refl} & \text{-} \\ \texttt{ctrsubj} & \text{suj} \\ \texttt{imp} & \text{-} \end{bmatrix}$$

Figure 1: Hypertag for *promettre* 'promise'

4. EASy Evaluation

Our first evaluation was conducted on the EasyDev corpus, a small corpus of around 4000 sentences used during the first EASy French parsing evaluation campaign and covering various document styles (journalistic, literacy, medical, mail, speech, etc.). The corpus is annotated with the EASy format (Paroubek et al., 2006; Paroubek et al., 2009), a mix of 6 kinds of chunks and 14 kinds of dependencies between forms or chunks, as illustrated by Figure 2 (with ovals for chunks and diamonds for dependencies). Table 3 shows the performances of the various lexica, on this EasyDev corpus. The coverage column indicates the rate of full parses (keeping in mind that the almost all remaining sentences get partial parses), and shows a clear decrease for DICOVALENCE and smaller ones for *LGLex* and NewL*efff*. We retrieve similar results in terms of F-measure on the chunks and dependencies. Finally, the fact that *LGLex* is both a wide-coverage and very fine-grained lexicon has a clear impact on parsing time. Figure 3 shows the F-measure for some of the EASy verbal dependencies, namely `SUJ-V` for the subject-verb relation, `AUX-V` for the auxiliary-verb relation, `COD-V` for the object-verb relation, `CPL-V` for the complement-verb relation (with no distinction between argument and adjuncts), and `ATB-SO` for the subject or object attributes. We have also added `MOD-N` for noun-modifiers, this relation being the most numerous one and being partly related to verbs through past and present participles on nouns.

Again, we observe a slight decrease for all new lexica versus L*efff*, more marked on some relations like `COD-V` for *LGLex* and `ATB-SO` for *LGLex* and DICOVALENCE.

Table 3 and Figure 3 also provide data for the first tried Alexina version of DICOVALENCE (dubbed "Old DICOVALENCE"). We can observe a much weaker coverage and very poor performances for the `ATB-SO` relation. Because of these figures, we were led to investigate and correct the conversion script that generates the Alexina version of DICOVALENCE, resulting in much better results parsing results.

Lexicon	**Cover. (%)**	**Chunks (%)**	**Rels (%)**	**Time (s)**
L*efff*	83.45	89.03	66.76	0.35
NewL*efff*	82.19	88.74	66.09	0.55
LGLex	80.61	87.89	63.19	1.10
DICOVALENCE	71.44	88.08	64.49	0.38
Old DICOVALENCE	65.69	87.06	62.72	0.42

Table 3: Overall performances on EasyDev

5. CoNLL Evaluation

For our second evaluation, we used the version of the French TreeBank (journalistic style) (Abeillé et al., 2003) converted by Candito et al. (2010) into the CoNLL dependency format, a format now largely used in international parsing evaluation campaigns (Nivre et al., 2007). This version of the French TreeBank has already been used to train and compare several statistical parsers (Candito et al., 2010), thus providing us baselines to evaluate FRMG and the various lexica. Note however that our results are still preliminary.

The CoNLL format relies on a fine-grained set of verbal dependencies, with in particular for verbal dependencies:

- the distinction between several kinds of auxiliaries: `aux-tps` (temporal auxiliaries), `aux-pass` (passive constructions), `aux-caus` (causative constructions);

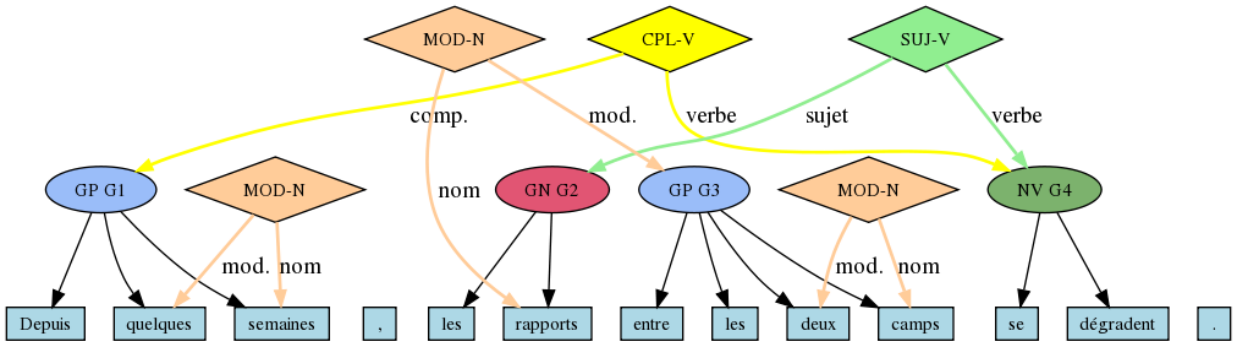

Figure 2: Sentence sample in Passage format 'Since a few weeks, relations between the two sides are deteriorating.'

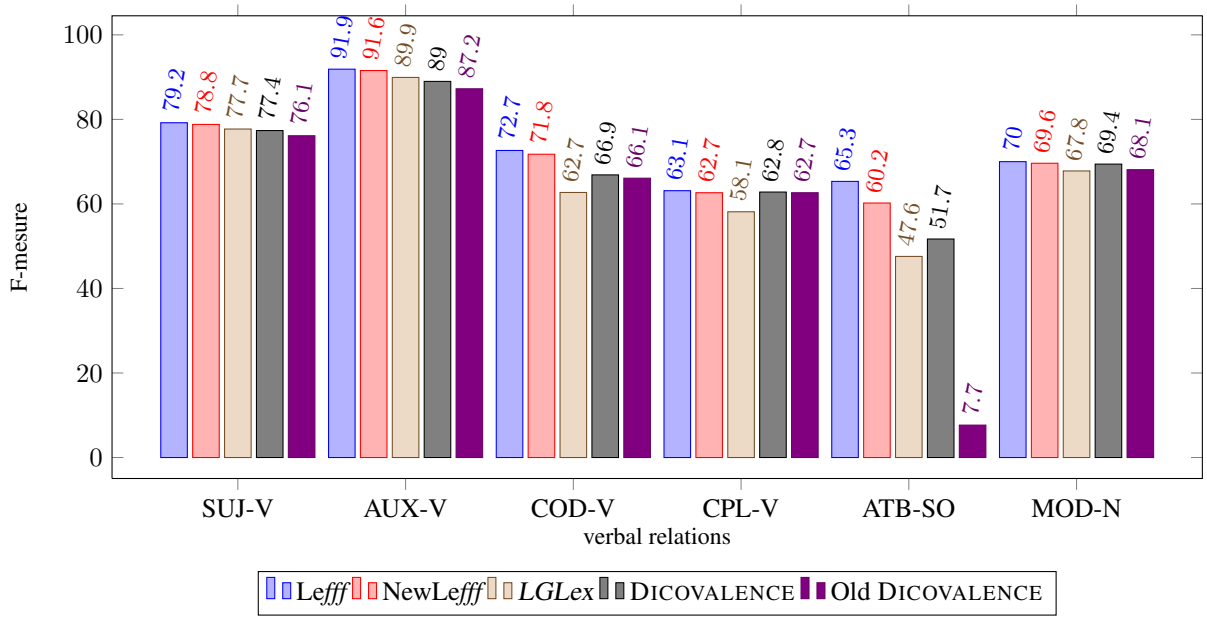

Figure 3: F-mesures for some verbal EASy relations

- `suj` and `obj` for the subjects and objects (but noting that the **obj** relation is also used in non-verbal cases);
- relations for the prepositional objects, with `a_obj` for those introduced by *à* (*to*), `de_obj` for those introduced by *de* (*of*), and `p_obj` for the remaining ones;
- relations for the attributes, with `ato` for the attributes of objects and `ats` for those of subjects;
- `aff` for affixes, actually verb clitics not covered by the above-mentioned relations;
- `mod` for verb modifiers such as adverbs (noting that again this relation is also used for non verbal cases)

Table 4 shows that all lexica got very good full parse coverage (on journalistic style) and emphasizes again the speed problems for *LGLex* (on relatively long and complex sentences, with a mean length average of 27 words vs 19.3 for EasyDev). Again, we note a slight decrease for the alternative lexica in terms of *Labeled Attachment Score* (LAS). We also note that all FRMG versions are still a few points below state-of-the-art statistical parsers, for instance MST (Candito et al., 2010). A finer analysis in terms of recall and precision at the dependency level shows a contrasted landscape (Figure 4), with *LGLex* or more specifically NewLe*fff* being sometimes better either in recall (`aff`, `a_obj`) or precision (`ato`, `aux_caus`). More generally, recall is relatively good but we observe precision problems. We conjecture that the finer granularity of *LGLex* and also of NewLe*fff* tends to wrongly select rare valency frames for some medium to high frequency verbs, frames that are strongly favored by the heuristic-based FRMG disambiguation algorithm, leading to confusion between verb arguments (`obj`, `a_obj`, `de_obj`, `p_obj`) and modifiers (`mod`, `dep`).

Lexicon	**Cover. (%)**	**LAS (%)**	**Time (s)**
Le*fff*	89.53	82.21	0.61
NewLe*fff*	88.76	81.36	0.94
LGLex	86.73	78.75	1.95
DICOVALENCE	75.28	79.38	0.69
MST	-	88.20	-

Table 4: Overall evaluation on French Tree Bank (test part)

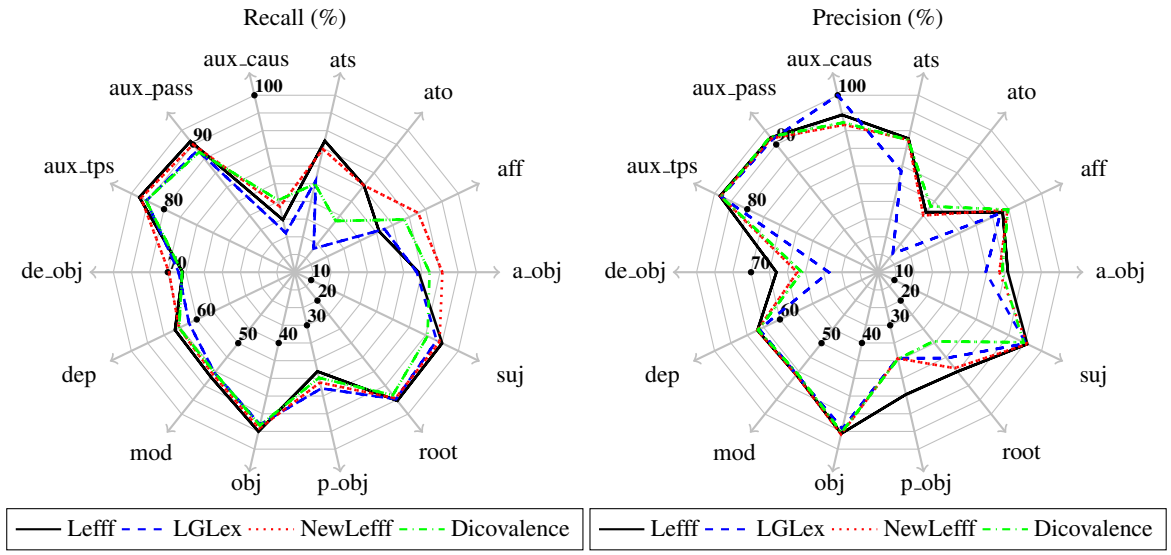

Figure 4: CONLL verbal relations

6. Error Mining

The evaluation already provides interesting feedback to identify the strong and weak points of a lexicon (as illustrated by DICOVALENCE with the ATB−SO relation). However, to get information at the level of a verb or of an entry, we rely on error mining techniques (Sagot and Villemonte de La Clergerie, 2006). More precisely, the basic idea is to identify *suspect* lexical entries by mining the full parse failures on a large corpus, based on the following intuition:

> A form is suspect if it occurs more often than expected in non-full-parsable sentences, in co-occurrence with non-suspect forms.

The mathematical formulation of this intuition leads to a fix-point iterative algorithm, close to EM (Expectation-Maximization), which may be used to also return, for each suspect w, a set of sentences in which w is suspected to be the cause of failure for a full parse. Suspects, their lexical entries, as well as associated sentences may then be browsed in a web interface to quickly identify the errors or lacks in the lexical entries.

This idea and its implementation may be directly used to track the errors in any new lexicon L. However, it is also interesting to contrast L with Le*fff*, considered as reference, which may be achieved through a slight rephrasing, as follows:

> A verb is suspect for lexicon L if it occurs more often than expected in sentences not full-parsable when using L but that received a full parse when using Le*fff*, in co-occurrence with non-suspect verbs.

The modified algorithm was then used on a larger corpus of 100K sentences (1.6M words), named CPJ (*Corpus Passage Jouet*), and comprising various style of documents (encyclopedic with Wikipedia, literacy with Wikisource, news with AFP, and discourse with Europarl). We have started exploiting the results for (former versions of) *LGLex* and, to a lesser extent, NewLe*fff*, and have identified several kinds of errors for some entries. The interest of this approach is that it can be applied to very large corpora, and we plan to do it, to overcome lexical data sparseness. Here, we provide some analysis of the data provided by the algorithm, with an emphasis on *LGLex*.

6.1. LGLex

We analyzed the first 15 suspicious verbs in *LGLex* in order to determine where the errors come from. We indicate the number of failed sentences for each verb between parentheses and we give one example. In total, there are 212 failed sentences for this selection of verbs:

- Some entries do not appear in the tables:
  - *mixer* 'mix' (7) in *Mixé par Jimi Hazel, assisté de Bruce Calder, enregistré chez Jimi à l' "Electric Lady Studios" à New York*: this entry is encoded in table **36S** but with the meaning 'blend' (*Max mixe les carottes (et+avec) les navets dans un mixeur*). We added this entry to table **32PL** (*Max a mixé les sons*), which has the defining feature <N0 V N1>, with an encoding similar to the entry *mélanger* 'blend' (*Max a mélangé les (étiquettes+cartes+couleurs)*);
  - *zapper* 'omit' (4) in *Elle a également "déploré" la mémoire de "plus en plus sélective" de la jeune femme, "qui zappe les détails qui font désordre"*: this entry appear in table **35L** but with the meaning 'channel hop' (*Max zappe de la 1ère et la 2ème chaîne*). We added this entry to table **32R2** (*Max a zappé un (repas+paragraphe)*), which has the defining feature <N0 V N1>, with an encoding similar to the entry *sauter* 'skip', 'miss' (*Max a sauté (un repas+une étape+une ligne)*);
  - *réaffirmer* 'reaffirm' (28) in *Nous réaffirmons la nécessité de consulter les sans-abri et leurs or-*

ganisations sur les programmes européens pertinents. We can add the feature <re−V> for the entry *affirmer* 'affirm' of table **9** (*Max a affirmé à Luc qu'il viendrait*)[10], with the defining feature <N0 V N1 à N1>, which accepts the construction <N0 V N1>, in order to recognize the entry *réaffirmer*;

- *réélire* 'reelect' (10) in *Helmut Kohl est réélu au poste de Chancelier fédéral*. We can add the feature <re−V> for the entry *élire* 'elect' of table **39** (*On a élu Mac (E+comme) député*), with the defining feature <N0 V N1 N2>, in order to recognize the entry *réélire*;
- the pronominal form *se réimplanter* 're-establish itself' (5) in *Celles-ci cherchent toujours à se réimplanter dans la zone", a relevé M. Besson*. The pronominal form *s'implanter* 'establish itself' does not appear either in the tables. We can add the feature <re−V> for the entry *implanter* 'establish' in table **38LD** (*On a implanté une usine dans cette région*), with the defining feature <N0 V N1 Loc N2 destination>, in order to recognize the entry *réimplanter* 're-establish'. Then, we can add the feature <se V> (or more precisely, <N1 se V W>) to accept the pronominal forms *s'implanter* and *se réimplanter* (*Une usine s'est implantée dans la zone*)[11].

- Some entries appear in the tables but are not encoded (∼) or have been corrected:
  - *susciter* 'spark off' (41) in *A d'autres niveaux, les propositions sur la table suscitent de sérieuses objections* and *recruter* 'recruit' (14) in *80 intérimaires ont déjà été recrutés, pour assurer les commandes*: these two entries appear in table **38R** (*Ceci a suscité une (vive réplique+réaction) chez Luc* and *Max a recruté Luc (comme+sur un poste de) lecteur*), which has the defining feature <N0 V N1 Loc N2>, but are not encoded. This implies that no other construction are accepted, whereas the construction <N0 V N1> appear in the table and allows the erasure of the second complement. We encoded this construction as +;
  - *réprouver* 'reprove', 'reprobate' (11) in *Dieu ne réprouve donc personne*: this entry is encoded in table **12** (*Max réprouve qu'Ida boive*), which has the defining feature <N0 V N1 de N2>. When we modified the defining features (Tolone, 2011), we replaced it by <N0 V N1>.

- Other entries are encoded in the tables but with obligatory complements which do not appear in the sentences of the corpus:
  - *délocaliser* 'relocate' (9) in *Ils ont fait le choix de délocaliser en Tunisie*: this entry appear in table **38L** (*On a délocalisé ce service de Paris à Dax*), which has the defining feature <N0 V N1 Loc N2 source Loc N3 destination>, but the entry is not encoded. We encoded the construction <N0 V N1 Loc N3 destination> as +, but no other construction allow the erasure of the first complement;
  - *kidnapper* 'kidnap' (12) and *revendre* 'retail' (5) in sentences without second complement, such as *Les deux Italiens ont été kidnappés le 18 décembre* and *Charles mangeait l'avoine des chevaux, doublant les fournitures, revendant par une porte de derrière ce qui entrait par la grande porte*: these two entries are encoded in table **36DT** (*On a kidnappé son fils à Max* and *Max a revendu à Luc la télé gagnée au jeu*), which has the defining feature <N0 V N1 Prep N2>, without allowing the erasure of the second complement introduced by the preposition *à*;
  - *écrouer* 'put behind bars' (5) in *Le lycéen de 18 ans soupçonné d'avoir poignardé vendredi un camarade, Hakim, dans leur lycée du Kremlin-Bicêtre (Val-de-Marne), a été mis en examen et écroué hier, alors que lycées et collèges sont invités à observer une minute de silence aujourd'hui à la mémoire de la victime*: this entry is encoded in table **38LHD** (*On a écroué Luc dans un pénitentier*), which has the defining feature <N0 V N1 Loc N2 destination>, without allowing the erasure of the locative complement;
  - *camper* 'camp' (5) in *Les troupes campent entre Harlem et Nimègue*: this entry is encoded in table **38LHR** (*Le roi campe ses troupes dans la plaine*), which has the defining feature <N0 V N1 Loc N2>, and it accepts the construction <N1 V W>. This means that an object N1 can become the subject of a sentence with conservation of its other objects. Therefore, this corresponds to the construction <N1 V Loc N2> (*Ses troupes campent dans la plaine*), without allowing the erasure of the locative complement.

- Last, some specific cases:
  - *rediriger* 'redirect' (50) in wrong sentences, such as *deux cent cinquante-troisredirige ici*;

[10] The entry *affirmer* appears also in table **32R3** *Max a affirmé sa (position+résolution)* but the difference is that it doesn't accept the intrinsic completive as here (*Paul a affirmé (la nécessité de+qu'il était nécessaire de) venir*).

[11] We could add the entry *s'implanter* to table **35L** but we don't because it shares exactly the same meaning with the entry *implanter* in table **38LD**. This case is different between the entry *fixer* 'screw' in table **38LD** (*Max a fixé le tableau au mur (E+avec des vis)*) and the entry *se fixer* 'settle' in table **35ST** (*Paul s'est fixé dans le midi*), which have different meanings. We don't add the entry *fixer* to table **38LD** with the meaning 'settle' because the transitive construction is not accepted (*?(On+ceci) a fixé Paul dans le midi*).

– *consoler* 'comfort' (6) in sentences with clitic pronominalization of the object, such as *Elle essayait de le consoler*: this entry is encoded in table **32R1** (*Max console le chagrin de Luc*), which has the defining feature <N0 V N1>, without allowing the clitic pronominalization of the first complement (we can add the feature like <N1 = Ppv =: le>).

The previous examples show (a) that some entries appear in the tables but are not encoded and therefore we have to encode them, and (b) that some other entries are missing, with several cases to be distinguished:

1. the entry should be added as a new entry. It can be a new verb or a different meaning of an existing verb (cf. *mixer* and *zapper*);

2. the entry *re-V* (or *ré-V*) has a meaning which can be derived by a "simple" use of the verb *V* (cf. *réaffirmer* and *réélire*): we should add a column <re-V> to all tables and encode it for all entries. Indeed, those without a derivable meaning have been added like another entry (entries *re-V* which they do not mean *faire une deuxième fois* 'do twice'), such as *revendre* 'retail' that does not mean *vendre une deuxième fois* 'sell twice' or *retomber* 'come down' (*La balle retombe*) which is not *tomber une deuxième fois* 'fall twice';

3. the entry *se V* has a meaning which can be derived by a transitive use of the verb *V* (cf. *s'implanter*): we have at least 5 different cases (Boons et al., 1976a) (p.120-163), so we should add 5 columns <se V> to all tables and encode it for all entries: for instance, *se regarder* 'look at oneself' (*Paul se regarde dans la glace*), *se mentir* 'lie to one another' (*Paul et Marie se mentent*), *s'étonner* 'be surprised' (*Paul s'étonne de mon silence*), *se laver* 'wash' (*Paul se lave les pieds*), *se manger* 'eat' 'be served' (*Le roti se mange froid*). Indeed, only the intrinsic pronominals, which are not linked by a transitive use, have been added as a new entry: for instance, *s'évanouir* 'faint' (*Paul s'est évanoui*);

4. the entry *se V* and *re-V* is a combination of the two previous cases (cf. *se réimplanter*);

5. the entry *dé-V* has a meaning which can be derived by a verb *V* (no example in this selection of verbs): we should add a column <dé-V> to all tables and encode it for all entries. Indeed, only the uses different from *faire l'action inverse* 'do the opposite action' have been added like another entry. For instance, *dévisser* has an entry for the meaning 'fall' (*L'alpiniste a dévissé*), but not for the meaning 'unscrew' (*dévisser une vis*), which is the opposite action of 'screw on' (*visser une vis*).

As we show, some features are also missing, including some that encode the erasure or the clitic pronominalization of certain complements. Indeed, we can allow as erasures the complements that are in the defining features, and we should add these features to the appropriate tables and encode them for all entries in all tables. For instance, the entry *donner* 'give' in table **36DT** can accept the erasure of all complements if the context allows it:

- <N0 V N1 Prep N2>: *Paul donne du fric aux associations*
- <N0 V N1>: *Paul donne du fric*
- <N0 V Prep N2>: *Paul donne aux associations*
- <N0 V>: *Paul donne souvent*

In conclusion, error mining enables us to detect and correct many errors in *LGLex* but we should manually analyze all 613 suspicious verbs in all 2,623 concerned sentences and all corrections have to be done manually, which represents an important (but rewarding) effort.

6.2. NewLe*ffff*

Error mining results on parses produced by FRMG with NewLe*ffff* on the large CPJ corpus were also investigated, although less thoroughly than for *LGLex*. It turned out that one verbal lemma was ranked by far the highest among all dubious lemmas, namely *estimer* 'consider'. Out of 569 sentences containing a form of this verb, as many as 200 could not receive a full parse. By looking at some of these failed sentences, we were able to quickly identify the following problem: the entry for *estimer* 'consider'[12] lacked clausal realizations for the direct object (finite clause and infinitive clause).

We also spotted a few other errors concerning *s'attendre à* 'expect' (missing control information), the attributive entry for *savoir* 'know' (missing clausal realizations for the object), *inciter* 'encourage, prompt' (missing infinitive realization for the subject), *réitérer* when meaning 'repeat' (missing clausal realization for the object), *se résoudre à* 'resolve to' (missing clitic and finite clause realizations for the indirect object), and others.

7. Conclusion

We have presented some preliminary but promising evaluation results for several lexica, obtained through their integration within a lexicalized deep TAG parser. Clearly, even if good, the results show that some efforts of adaptation remain to be done to improve the integration and to better exploit the richness of these lexica. Error mining techniques should help us to achieve this objective, and should also help us to identify the strong and weak points of each lexicon, which should lead to a new generation of better quality lexica, freely available and ready to use in large scale NLP systems.

We would like also to mention very recent results showing that partially supervised learning techniques may be used to boost the performance of FRMG disambiguation to reach a LAS of 85.1% when using Le*fff*, to be compared with the 82.2% presented in this paper — and much closer to MST, a stochastic parser specifically trained on the French Treebank. It remains to be tested whether this improved disambiguator leads as such to similar gains when using the other lexica, or whether the learning phase has to be done for each of them.

[12] In NewLe*fff*, *estimer* has three entries, that corresponds to the meanings 'consider', 'estimate' and 'esteem'.

8. References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks*. Kluwer, Dordrecht.

Claire Blanche-Benveniste, José Delofeu, Jean Stefanini, and Karel van den Eynde. 1984. *Pronom et syntaxe. L'approche pronominale et son application au français*. SELAF, Paris.

Jean-Paul Boons, Alain Guillet, and Christian Leclère. 1976a. *La structure des phrases simples en français : Constructions intransitives*. Droz, Genève, Suisse.

Jean-Paul Boons, Alain Guillet, and Christian Leclère. 1976b. La structure des phrases simples en français, classes de constructions transitives. Technical report, LADL, CNRS, Paris 7.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, Valletta, Malta.

Matthieu Constant and Elsa Tolone. 2010. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Michele De Gioia, editor, *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008), Seconde partie*, volume 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, pages 79–193. Aracne, Rome, Italie.

Éric de La Clergerie. 2005. From metagrammars to factorized TAG/TIG parsers. In *Proceedings of IWPT'05 (poster)*, pages 190–191, Vancouver, Canada.

Maurice Gross. 1975. *Méthodes en syntaxe : Régimes des constructions complétives*. Hermann, Paris, France.

Aravind K. Joshi, Leon Levy, and Makoto Takahashi. 1975. Tree Adjunct Grammars. *Journal of Computer and System Science 10*, 10(1):136–163.

Alexandra Kinyon. 2000. Hypertags. In *Proc. of COLING*, pages 446–452.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *The CoNLL 2007 shared task on dependency parsing*.

Patrick Paroubek, Isabelle Robba, Anne Vilnat, and Christelle Ayache. 2006. Data, Annotations and Measures in EASy, the Evaluation Campaign for Parsers of French. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC'06)*, Gênes, Italie.

Patrick Paroubek, Éric Villemonte de la Clergerie, Sylvain Loiseau, Anne Vilnat, and Gil Francopoulo. 2009. The PASSAGE syntactic representation. In *7th International Workshop on Treebanks and Linguistic Theories (TLT7)*, Groningen, January.

Benoît Sagot and Laurence Danlos. 2012. Merging syntactic lexica: the case for French verbs. In *Proceedings of the LREC 2012 workshop on Merging Language Resources*, Istanbul, Turkey. Submitted.

Benoît Sagot and Karën Fort. 2009. Description et analyse des verbes désadjectivaux et dénominaux en *-ifier* et *-iser*. *Arena Romanistica, Journal of Romance Studies*, 4:102–110. ISSN 1473-3536.

Benoît Sagot and Éric Villemonte de La Clergerie. 2006. Error mining in parsing results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 329–336, Sydney, Australia, July.

Benoît Sagot. 2010. The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, Valletta, Malta.

Elsa Tolone and Benoît Sagot. 2011. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. In Zygmunt Vetulani, editor, *Human Language Technology, Forth Language and Technology Conference, LTC 2009, Poznań, Poland, November 2009, Revised Selected Papers*, Lecture Notes in Artificial Intelligence (LNAI). Springer Verlag.

Elsa Tolone, Éric de La Clergerie, and Benoît Sagot. 2011. Évaluation de lexiques syntaxiques par leur intégration dans l'analyseur syntaxique FRMG. In *Colloque Lexique et grammaire 2011*, Nicosia, Cyprus, October.

Elsa Tolone. 2011. *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. Ph.D. thesis, LIGM, Université Paris-Est, France. (340 pp.).

Karel van den Eynde and Piet Mertens. 2006. Le dictionnaire de valence DICOVALENCE : manuel d'utilisation. http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.