Mapping the Lexique des Verbes du Français (Lexicon of French Verbs) to a NLP Lexicon using Examples

Bruno Guillaume, Karën Fort, Guy Perrier, Paul Bédaride
Inria/LORIA, Université de Lorraine/LORIA
Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France
firstname.lastname@loria.fr

Abstract

This article presents experiments aiming at mapping the Lexique des Verbes du Français (Lexicon of French Verbs) to FRILEX, a Natural Language Processing (NLP) lexicon based on DICOVALENCE. The two resources (Lexicon of French Verbs and DICOVALENCE) were built by linguists, based on very different theories, which makes a direct mapping nearly impossible. We chose to use the examples provided in one of the resource to find implicit links between the two and make them explicit.

Keywords: language resources mapping, syntactic lexicon, semantic lexicon

1. Introduction

Despite continuous efforts from some members of the community, Natural Language Processing (NLP) still needs more quality linguistic resources for French. One solution consists in adapting existing resources, often created by linguists by hand. These resources were not intended to be used by machines and they have to be – sometimes heavily and always imperfectly – formatted, therefore losing information in the process.

The LEXIQUE DES VERBES DU FRANÇAIS (LVF, Lexicon of French Verbs) (Dubois and Dubois-Charlier, 1997) is no exception to that rule and several research works showed both the interest and the difficulty there is in adapting this resource for NLP (Hadouche and Lapalme, 2010; Bédaride, 2012). Our work is in line with this research and we present here the – still incomplete – results of our efforts in adapting the LVF for NLP.

Our primary objective is to improve the coverage of FRILEX, a morphological and syntactic lexicon we use for parsing using the tool LEOPAR (Perrier and Guillaume, 2013) and to benefit from semantic information provided by LVF.

FRILEX was partly built from DICOVALENCE (Mertens, 2010), which coverage is relatively limited (3,729 lemmas in DICOVALENCE, versus 12,308 lemmas in LVF). The advantage of DICOVALENCE with respect to lexicons with a larger coverage, like Le*fff* (Sagot, 2010), is that it includes rich and fine information (semantical restriction, passive reformulation, cross-references between entries in case of reformulation...), which is completely validated by linguists. DICOVALENCE is based on a linguistic theory that is very different from that of LVF, which makes the mapping between the two resources very difficult, if not impossible. However, we hypothesized that we could use the examples provided in the LVF to help aligning, at least partly, the two lexicons.

This paper is organized in the following way: in section 2. we present the paradigms that led to the construction of the LVF and DICOVALENCE, then we detail in section 3. the methodology we used to unfold the LVF examples and finally, we present in section 4. the obtained mapping between lexicons.

2. The LVF and DICOVALENCE: Two Very Different Construction Paradigms

2.1. DICOVALENCE

The originality of the verb lexicon DICOVALENCE (Mertens, 2010) is that it is based on the pronominal approach. The principle of the pronominal approach were introduced by Karel Van den Eynde et Claire Blanche-Benveniste (van den Eynde and Blanche-Benveniste, 1978; Blanche-Benveniste et al., 1984). In this setting, the authors use the term "valency frame" instead of "subcategorization frame". The valency frame is described using paradigms that correspond to the syntactic arguments governed by the verb; each paradigm is described by the set of pronouns which can syntactically fill the paradigm. Despite the focus on syntax, DICOVALENCE contains also semantic information: with each entry, a Dutch translation of the French verb is given. A more recent version (2010) also includes English translations. Figure 1 presents an entry of DICOVALENCE (version 1.2) for a sense of the verb *arrêter* (to stop).

The core of the above entry is composed of the P0, P1 and PP fields, which respectively represent the subject, the direct object and the manner complement of *arrêter*, that is the three elements of its valency frame. An entry may also contain additional syntactic information. In this example, the LC field presents the three possible diathesis changes for this verb.

It has to be noted that though the coverage of this lexicon is quite limited (it covers 8,334 senses of 3,729 verbs), it corresponds, according to its authors, to the most frequently used verbs in French.

2.2. The LVF(s)

The LVF (Dubois and Dubois-Charlier, 1997) was created by two French linguists, Jean Dubois and Françoise Dubois-Charlier. Their goal was to provide a linguistic description for French verbs, based on the idea, from Levin (1993), that the subcategorization frames of the verbs are

```
VAL   : arrêter: P0 P1 (PP<avec>)
VTYPE : predicator simple
VERB  : ARRETER/arrêter
NUM   : 7270
EG    : ils ont arrêté les travaux après l'accident
TR_DU : tegenhouden, ophouden, stilzetten, stilleggen, tot staan brengen
TR_EN : stop
P0    : qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
P1    : que,qui,te,vous,la,le,les,se réc.,en Q,ça,ceci,celui-ci,ceux-ci,l'un l'autre
PP_PR : avec
PP    : 0, quoi, ceci, celui-ci, ceux-ci
RP    : passif être, se passif, se faire passif
LC    : 7270-7280 je l'arrête avec ça, ça l'arrête
AUX   : avoir
```

Figure 1: DICOVALENCE entry for a sense of *arrêter* (to stop)

linked to their semantic interpretation. The lexicon contains 12,308 lemmas, corresponding to 25,609 senses. The lemmas are distributed into levels, corresponding to the frequency of usage of the verb (level 1 comprising the most frequently used verbs).

After years of unavailability, the lexicon was finally publicly released in 2007, in the form of an MS Excel file (ELVF) and a growing number of researchers in NLP got interested in it. However, as the lexicon was built manually, its usage with computers soon proved to be complex. In particular, a number of words were truncated or abbreviated and its representation as a table limits the richness of the structure. Finally, the codes and formats used in this original file cannot be understood without referring to the book (Dubois and Dubois-Charlier, 1997).

A XML version of the LVF (XLVF) was developed (Hadouche and Lapalme, 2010), making it more accessible, usable and extensible. To achieve this, the description of the codes used in the ELVF was encoded into XML files. Then, a XML file with (some) uncompressed data from ELVF was generated. For example, for the verb "*amasser*" (to stock), the "1aZ" code was uncompressed and generated the following XML snippet:

```
<conjugaison
  aux="avoir (sauf si pronominal ou
              entrée en être)"
  groupe="1"
  sous-groupe="chanter">
    1aZ
</conjugaison>
```

An extended version of XLVF, EXLVF was then created (Bédaride, 2012), further refining the resource. The most significant improvement that it provides is to decompose the OPÉRATEUR (operator) field and to link it to the CONSTRUCTIONS fields, in order to be able to associate different syntactic realizations with one semantic meaning.

Another important achievement was to provide an automatic parsing of the example sentences, as well as a dependency structure for each of them, linked to a subcategorization frame.

However, using examples from the LVF is still not trivial, as they are not directly usable:

- verbal forms are replaced by their first letter followed by the $\sim$ symbol, as in: *"L'avion v$\sim$ de Paris"* (The plane a$\sim$ from Paris) instead of *"L'avion vient de Paris"* (The plane arrives from Paris).
- several examples are written as one sentence in a factorized way, as in: *"On é$\sim$ le concert,ce musicien à cinq heures."* (We listened to this concert, this musician at five) instead of *"On é$\sim$ le concert à cinq heures."* (We listened to this concert at five) and *"On é$\sim$ ce musicien à cinq heures."* (We listened to this musician at five)

More examples of these problems are given in figure 2.

In EXLVF, most of the factorized examples were unsplit using transformation rules. To deal with the elided verbal forms, EXLVF was created using a statistical parser trained on a modified version of the French TreeBank (Abeillé et al., 2003), where verbal forms were replaced by elided ones.

3. Unfolding the LVF Examples

Figure 2 gives several examples of phenomena involved in the unfolding process (numbers refer to sentences in the tables).

We concentrated first on unfolding the examples provided in the lexicon, that is:

1. de-factorizing the sentences (1 $\Rightarrow$ 3, 4), and
2. rebuilding the verbal forms (3 $\Rightarrow$ 6).

3.1. Procedure

In EXLVF, the de-factorization part of the process was partly conducted by an automatic transformation. Unfortunately, the factorization of the examples does not follow any regular template and the automatic transformation produces some ungrammatical sentences. However, we used it as the starting point for our process.

In this work, we used the LEOPAR parser (Perrier and Guillaume, 2013), a symbolic parser based on Interaction

LVF, arrêter (02)	exLVF		LEOPAR
1: *On a∼ la télé,l'aspirateur. (We switched off the TV, the vacuum cleaner.)*	3: *On a∼ la télé.*	=	6: *On arrête la télé.*
	4: *On a∼ l'aspirateur.*	=	7: *On arrête l'aspirateur.*
2: *La machine s'a∼. (The machine stops)*	5: *La machine s'a∼.*	≠	8: *La machine s'arrête.*

LVF, arrêter (05)	exLVF	LEOPAR	Manual	LEOPAR
9: *On a∼ de travailler, de fumer, de courir, le sport. (We stopped working, smoking, running, practicing)*	10: *On a∼ de travailler, le sport.*	Fail	13: *On arrête de travailler.*	Ok
	11: *On a∼ de fumer, le sport.*	Fail	14: *On arrête de fumer.*	Ok
	12: *On a∼ de courir, le sport.*	Fail	15: *On arrête de courir.*	Ok
			16: *On arrête le sport.*	Ok

Figure 2: Examples of the unfolding operation on the LVF examples for "*arrêter*" (to stop)

Grammars. Being symbolic, it is able to detect ungrammatical sentences and to reject them. We used this feature in both subtasks: first, to detect sentences that were incorrectly unsplit in exLVF and, second, to recover the correct verbal form that is elided in exLVF.

We proceeded in four steps:

1. Parsing of the examples from exLVF with LEOPAR: the elided form *"x∼"* is replaced by a sequence of possible forms for the given verb lemma; the first form with which the sentence can be parsed is considered as the correct one.
2. Comparison of the results with those from the statistical parsing of exLVF (for example, same analyses for 3-6 and for 4-7, different analyses for 5-8).
3. Manual unfolding or correction of the examples that were left unparsed by LEOPAR (10, 11, 12 ⇒ 13, 14, 15, 16).
4. Parsing with LEOPAR of the manually checked examples.

The parsing of exLVF left 2,107 sentences unparsed.

3.2. Results

We processed all the DICOVALENCE lemmas with the method described above: 30.2% of the lemmas from the LVF, out of which 835 of the 850 level 1 verbs from LVF and 2,937 out of the 3,096 level 2 verbs[1]. This coverage corresponds to 46.8% of the senses (entries) of the LVF lexicon.
In the first step, 26,298 examples were given to LEOPAR. Among them, 24,191 were successfully parsed.
In the second step, we compared ours with the exLVF analyses and found 11,379 occurrences where parsing results were identical and 12,812 occurrences where they differed.
In the third step, the 2,107 unparsed sentences were unfolded manually. This was done by the authors without overlapping; indeed, in LVF, examples are written for a human reader and in most of the cases, unsplitting them is straightforward and not difficult (we found only a few cases where the unfolding was ambiguous). This unfolding phase produced 2,626 sentences.
In the last step, 1,059 of the 2,626 sentences were successfully parsed with LEOPAR, leaving 1,351 sentences unanalyzed. In this last case, it is possible to recover the parsing given by exLVF on the sentence before manual splitting as a backoff.
The sentences that remain unanalyzed generally correspond to sub-categorization frames from LVF that are unknown to DICOVALENCE.
This procedure allowed us to provide, for each example we analyzed, a syntactic analysis associated with reliability indicators:

[C] if exLVF and LEOPAR provided the same analysis (11,379 occurrences);

[D] if exLVF and LEOPAR provided two different analyses (12,812 occurrences);

[L] if the example was manually split or corrected and an analysis was produced by LEOPAR (1,095 occurrences);

[Z] if the example was manually split and no analysis was produced (1,531 occurrences).

3.3. Evaluation

We performed a manual evaluation on 100 examples that are similarly analyzed by both systems, in order to check if the parsing was actually correct. Out of the 100 examples, 92 are correctly analyzed.
We also manually checked 100 examples that are analyzed differently by the tools: 41 of them are correctly parsed by LEOPAR, 30 by exLVF and the remaining 29 are wrongly analyzed by both tools.
It has to be noted that this manual evaluation was done without overlapping (only one expert did the checking) by three persons that can be considered experts.

4. Tracing back the Lexicons

As we already mentioned, DICOVALENCE and LVF originated from very different linguistic theories, that imply different views on the senses of a verb, hence different ways of dividing them up. As an example, there are 18 senses for "*compter*" (to count) in LVF, while DICOVALENCE enumerates 17 entries for the verb.[2]

[1]For more details, see: `http://wikilligramme.loria.fr/doku.php?id=lvf`

[2]Note that "*compter*" is one of the most polysemous verb in DICOVALENCE.

4.1. Procedure

In order to merge, compare, complete and make interoperable both resources, it is essential to correctly align the senses between them and to build the relation that allows to map them. There is no unequivocal relation between the two resources and one entry from one of them can correspond to any number of entries (including none) from the other one.

Building this alignment requires significant manual work, which can only be done precisely by an expert who has a good knowledge of both resources. We propose here to carry out part of the work automatically, analyzing the examples from one lexicon using the other.

We applied this idea analyzing the examples from LVF with the FRILEX lexicon, built on DICOVALENCE. In FRILEX, each entry describing a sub-categorization frame is associated to the list of DICOVALENCE entries that covers this frame. We built a binary relation between LVF entries and DICOVALENCE entries: a LVF entry L is associated to a DICOVALENCE entry D if LEOPAR parses one of the example associated to L using a FRILEX entry built from D. In this case, we said that L and D are *syntactically compatible*.

4.2. Results

Figure 3 shows the relations we obtain with "*compter*" (to count).

In order to make the figure easier to read, LVF senses and DICOVALENCE entries are clustered with respect to equivalent classes in the mapping. Thus, *compter*_4, as in "On c$\sim$ les taxes dans le prix." (taxes are included in the price), is syntactically compatible with the three entries numbered 17,380, 17,415 and 17,425.

4.3. Evaluation

In order to evaluate our results, we had an expert build a gold standard mapping for the verb "*compter*": for each of the 18 LVF senses, he chose the corresponding DICOVALENCE entry.

For 14 senses, the correct DICOVALENCE entry is one of the entries given by our procedure. For 3 senses, the correct DICOVALENCE entry in not covered by the conversion from DICOVALENCE to FRILEX. For the last sense, LEOPAR was wrong (it found the right mapping but with a parsing solution which is not ranked first).

We also performed a global evaluation and observed the general ambiguity of the mappings: for 1 LVF sense, how many DICOVALENCE entries are selected? Figure 4 shows the results we obtain.

The red bars correspond to the baseline where all LVF senses for a lemma are mapped to all DICOVALENCE entries for this lemma. The green bars correspond to our mapping. Each bar indicates the number of LVF senses that are mapped to a given number of DICOVALENCE senses: for instance the higher bars (obtained for one DICOVALENCE entry) reflects the fact that before filtering 3,094 LVF senses (red) were unambiguously map to only one DICOVALENCE entry; after filtering (green), this is the case for 5,796 senses.

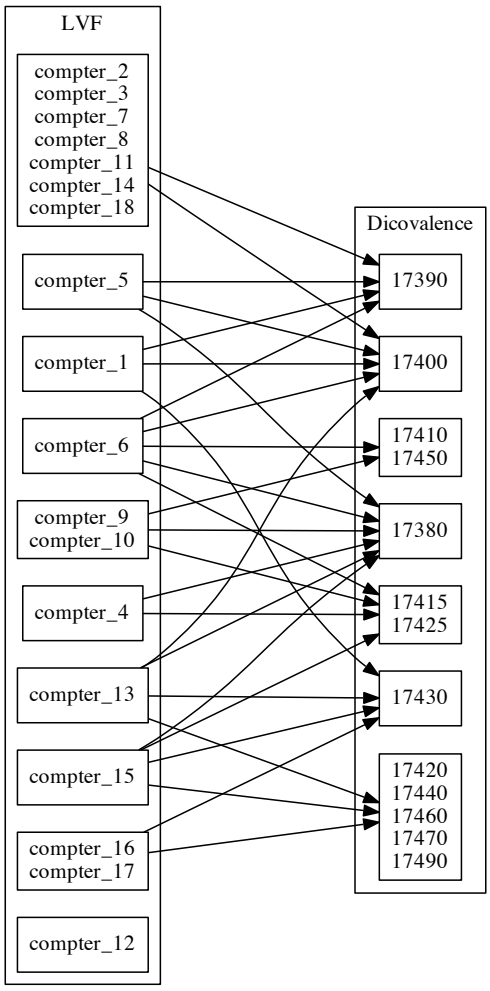

Figure 3: Mapping LVF - DICOVALENCE for "*compter*" (to count)

Of course, our goal is to have as much as possible mapping of LVF senses to a small number of DICOVALENCE entries (*i.e.* to have higher bars around the small number of DICOVALENCE entries). We observe that the ambiguity is much lower: without taking into account senses with 0 output, the mean ambiguity is 3.99 for the baseline and 1.95 for our mapping.

Conclusion

We showed in this article that it is partly possible to semi-automatically map one syntactic lexicon to another one, even when they are based on very different linguistic theories, thanks to the examples they provide.

For our first experiment, LVF examples were parsed with a DICOVALENCE-based lexicon. We chose this setting because the DICOVALENCE-based FRILEX lexicon already exists, but we can of course imagine to apply the same methodology in both ways to any pair of syntactic lexicons that contains linguistic examples.

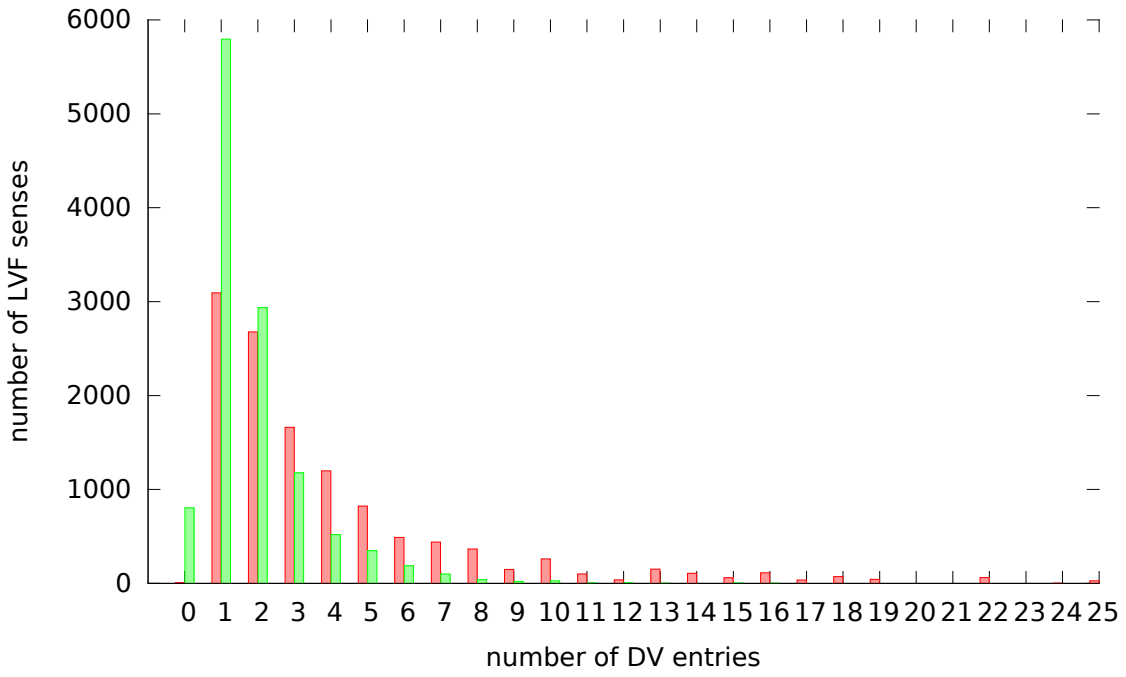

Figure 4: Ambiguity in the mapping LVF - DICOVALENCE before (in red) and after (in green) our filtering

Apart from the two lexicons described here, the Lexique-Grammaire (Gross, 1975) is certainly one of the resources that can be used with this method. For syntactic lexicons without examples (like Le*fff* (Sagot, 2010)), it is still possible to build such a mapping (only in one way, of course) with a lexicon containing examples.

5. References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In Abeillé, A., editor, *Treebanks*, pages 165 –187. Kluwer, Dordrecht.

Bédaride, P. (2012). Raffinement du lexique des verbes français. In *Proceedings of the Conference Traitement Automatique des Langues Naturelles (TALN)*, pages 155–168, Grenoble, France, June. ATALA/AFCP.

Blanche-Benveniste, C., Delofeu, J., Stefanini, J., and van den Eynde, K. (1984). *Pronom et syntaxe. L'approche pronominale et son application au français*. Sociolinguistique, systèmes de langues et interactions sociales et culturelles. SELAF.

Dubois, J. and Dubois-Charlier, F. (1997). *Les verbes français*. Larousse-Bordas, Paris, France.

Gross, M. (1975). *Méthodes en syntaxe*. Hermann, Paris, France.

Hadouche, F. and Lapalme, G. (2010). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages*, 10:193–220.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. University of Chicago Press.

Mertens, P. (2010). Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE Conversion vers un format utilisable en TAL. In *Proceedings of the Conference Traitement Automatique des Langues Naturelles (TALN)*, Montreal, Canada.

Perrier, G. and Guillaume, B. (2013). Leopar: an Interaction Grammar Parser. In *Proceedings of the Workshop on High-level Methodologies for Grammar Engineering, ESSLLI*, pages 121–122, Düsseldorf , Germany.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the international conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

van den Eynde, K. and Blanche-Benveniste, C. (1978). Syntaxe et mécanismes : présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32:3–37.