Adapting VerbNet to French using existing resources

Quentin Pradet, Laurence Danlos, Gaël de Chalendar

Adapting VerbNet to French using existing resources

Quentin Pradet[1], Laurence Danlos[2] and Gaël de Chalendar[1]
1 CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Gif-sur-Yvette, F-91191, France
2 Université Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA

Abstract

VerbNet is an English lexical resource for verbs that has proven useful for English NLP due to its high coverage and coherent classification. Such a resource doesn't exist for other languages, despite some (mostly automatic and unsupervised) attempts. We show how to semi-automatically adapt VerbNet using existing resources designed for different purposes. This study focuses on French and uses two French resources: a semantic lexicon (Les Verbes Français) and a syntactic lexicon (Lexique-Grammaire).

1 Introduction

Natural Language Processing needs lexicons and a large amount of data to efficiently analyze open-domain text. Getting this amount of data is a problem in itself, and is known as the knowledge acquisition bottleneck in the word sense disambiguation literature (Gale et al. 1992). While annotating more and more data will reduce the bottlenecks for some domains, encoding lexicons in a cost-effective way can lead to better improvements by explicitly stating similarities and differences between words.

The two main issues faced by sense lexicon creators are sense granularity and sense distinction, both of which are addressed by Levin classes (Levin 1993). In those classes, verbs are classified according to their semantics and to their syntactic alternations. VerbNet (Kipper-Schuler 2005) is an electronic lexicon which is inspired by Levin classes. It encodes thematic roles and semantic decomposition (Section 3). New constructions, new classes and new corrections have been added to VerbNet over the years.

With VerbNet, one can use a syntactic alternation to map syntactic chunks of a sentence to thematic roles (Swier and Stevenson 2005; Pradet et al. 2013). This task, semantic role labeling, has grown steadily in importance: it helps information extraction (Surdeanu et al. 2003), question-answering (Shen and Lapata 2007), event extraction (Exner and Nugues 2011), plagiarism detection (Osman et al. 2012), machine translation (Bazrafshan and Gildea 2013), or even stock prediction (Xie et al. 2013). Thanks to its high coverage (more than four thousand distinct verbs) and useful verb separation, VerbNet is well-suited for semantic role labeling.

However, a high-quality version of VerbNet is currently only available for the English language. Such a resource would be even more useful for less-resourced languages where role-labeled corpora are missing. VerbNet has cross-linguistic potential, as shown with the Portuguese language (Kipper-Schuler 2005, section 2.2.2). Adapting VerbNet to a new language would make it possible to reuse its structure, keep most semantic information, and produce a useful lexicon without years of manual work.

With the goal of developing a French version of VerbNet (called Verb∋net) in mind, we first have translated top-level VerbNet members in French and used French linguistic resources that encode the syntactic and semantic behavior of verbs (Section 3) to keep only the right translations (Section 4.1). The second step in building Verb∋net, which is still underway, is to adapt VerbNet syntactic alternations for French, which gives rise to various problems that we will discuss in Section 4.2.

2 Related work

Translating Levin classes and more recently VerbNet is recognized as a useful task in the literature. First, automatic methods have led to improvements in VerbNet itself: new classes have been incorporated (Korhonen and Briscoe 2004) and new verbs have been added from the LCS database (Dorr et al. 2001) or using the Sketch Engine tool (Bonial et al. 2013).

In other languages, Merlo et al. (2002) have used crosslinguistic similarities to convert 20 Levin classes to Italian. Automatic acquisitions have also been led in Japanese (Suzuki and Fukumoto 2009), German (Im Walde 2006), and Spanish (Ferrer 2004). The only direct translations we are aware of are the Estonian VerbNet (Jentson 2014) and the Brazilian Portuguese VerbNet (Scarton and Alusio 2012), which uses the mappings between VerbNet and WordNet, and between WordNet.Br and WordNet.

In French, Saint-Dizier (1996) first produced a VerbNet-like resource. To the best of our knowledge, effort on this resource has stopped and the result is not available. Later work has focused on automatic acquisition of subcategorization frames which were clustered according to their syntactic and semantic similarity. Sun et al. (2010) used a large subcategorization frame lexicon (Messiant et al. 2010) to cluster verbs according to two types of features: syntactic (subcategorization frames) and semantic (collocations and lexical preferences of verbs). Evaluation against a manually created gold standard led to an F-measure of 55.1%. Falk et al. (2012) apply a different clustering algorithm, and use different features, improving the F-measure to 70% on a similar but easier dataset. While those resources

highlight new ways to separate French verbs, the errors they contain are only one source of errors in applications: it is important to correct them if possible. While the results can still be improved, we believe that there is also a need for a manually validated French VerbNet. Our translation will be linked to the English VerbNet and the two linguistic resources we use, Les Verbes Français and the Lexique-Grammaire. It will also be open: we want to foster external contributions with our web-based tool and make the resource easy to use by using the same file formats than VerbNet.

3 Presentation of VerbNet and French Lexical resources

The top hierarchy in VerbNet is made up of 270 classes. Any class can be further divided into sub-classes organized in a tree structure. For each (sub-)class, this electronic lexicon gives: the list of verbs belonging to it, the relevant thematic roles which are possibly associated with selection restrictions, and the list of frames. A frame includes an illustrating example, a syntactic formula explicating the relation between syntactic arguments and thematic roles, and a semantic formula based on predicates that denote relations between participants and events.
Starting in the 70's two main lexical resources for French verbs, LVF and LG, were manually developed:

- LVF (Les Verbes Français, Dubois and Dubois-Charlier (1997)) includes around 25000 entries classified into 14 semantic classes with 54 syntactico-semantic subclasses and 248 subsubclasses.
- LG (Lexique-Grammaire, Gross (1975) and Boons et al. (1976)) includes around 14000 entries classified into 67 "tables", each table grouping verbs with the same frames and possibly with similar semantics. Each column of the table encodes additional restrictions that will apply to some of the verbs of that table.

Both LVF's classes and LG's tables can be compared to VerbNet's classes. However, these (old) French resources record neither thematic roles nor semantic formulae[1]. This is why we want to build a new French resource, Verb∋net. It will take advantage, on the one hand, of the existing French resources with a rich encoding of syntactic information, and on the other hand, of the semantic information in VerbNet built for English, a language relatively close to French.

4 Building Verb∋net

Our basic principle is that the top hierarchy in Verb∋net should be as close as possible to that in VerbNet with 270 classes. Nevertheless, some classes may disappear. This can be simply due to morphological reasons. Any VerbNet class made up only of verbs that are zero-related to nominals doesn't have a French equivalent, eg. class pit-10.7 with verbs such as *bark* and *bone* or week-end-56 with verbs such as *week-end* and *december*. On the other hand, class 10.8 with verbs formed by the prefix *dé-* plus a nominal (*debark, debone*) does have a French equivalent with verbs formed by the prefix *dé-* or *é-* (*déveiner, équeuter*).
Given this basic principle, building Verb∋net goes in two steps.

4.1 First step

The first step in building Verb∋net was to determine which French verbs belong to one of VerbNet's 270 classes. This was done in three stages:

1. For a given VerbNet class C_e, we manually assigned the LVF class(es) C_{lvf} and the LG's table(s) C_{lg} that fit its semantic definition (e.g. put-9.1 L3b 38LD or body_internal_motion-49 M1a 32CL or 32R3 or 32C),
2. we used two bilingual dictionaries (SCI-FRAN-EURADIC and the French Wiktionary) which give the list L_{trad} of the French translations of the English verbs belonging to C_e or a subclass of C_e,
3. we computed the verbs of the French class C_f which are a priori the simple verbs of L_{trad} which belong to the intersection of C_{lvf} and C_{lg} (e.g. *mettre, poser* or *installer* in put-9.1).

This step was performed quickly and gave accurate results: by keeping only verbs at the intersection of L_{trad}, C_{lvf} and C_{lg}[2], the results are precise and syntactically and semantically coherent. For example, the scribble-25.2 class contains 18 verbs in English; it is associated with LVF R3a.1 and LG 32A, which leads to a list of 16 French verbs: *composer*, *couper*, *donner*, *exécuter*, *fabriquer*, *faire*, *forger*, *former*, *imprimer*, *lever*, *produire*, *reproduire*, *sculpter*, *tailler*, *tirer* and *tracer*. All these verbs are valid for this class. This method results in a lexicon with 4058 verbs (2128 distinct verbs).

4.2 Second step

The second step in building Verb∋net has proven much more difficult than the first. For each of the 270 C_f sub-classes, we determine whenever possible:

- the possible subclasses in order to assign the verbs found in the first step to one of these sub-classes (if possible)
- the frames that are valid for French with possible adjustments for thematic roles and selection restrictions.

This step has first required to develop an editing tool (Section 4.2.1) to help and maintain the lexicographers' work. Next, it has required to set up basic

[1] The uses of thematic role and event were not much widespread in the 70's.

[2] When the intersection is empty, the non-empty list (C_{lvf} or C_{lg}) was chosen.

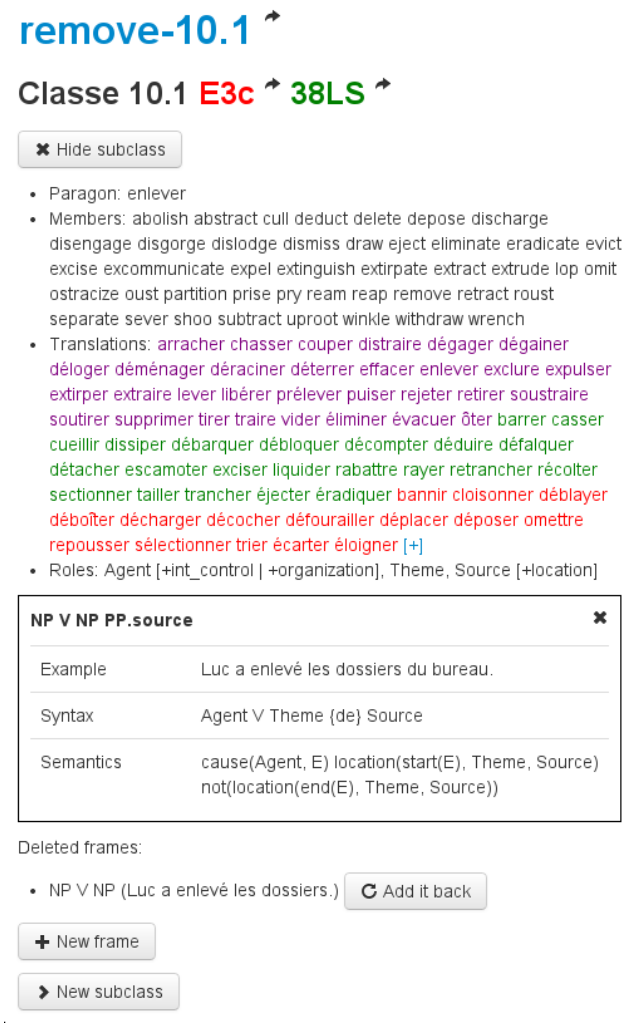

Figure 1: Web interface to analyze and edit Verb∋net. Every frame can be modified and the structure can be reorganized. The translations in purple belong to the intersection of C_{lvf} and C_{lg}; the translations in red (resp. green) belong only to C_{lvf} (resp. C_{lg}).

principles on French frames, when they differ from English ones (Section 4.2.2). Finally, a fine grained case-by-case study reveals some tough differences between French and English, which are illustrated in (Section 4.2.3).

4.2.1 Verb∋net editing tool

This step required us to develop a web-based tool which makes it possible to collaboratively edit VerbNet classes and frames by manipulating their representation on the website. This online interface developed with Django (a Python web framework) hides a PostgreSQL database that stores all this information and tracks all changes to the data. The tool was first filled with VerbNet frames and verb translations found in the first step. It allows us to edit a frame and to suppress or add a (sub-)class or a frame. For example, all the frames involving a conative, dative or benefactive alternation can be systematically suppressed because these alternations don't exist in French.
With the help of this tool (illustrated in Figure 1), the work for the second step can be very easy. For example, the four sub-classes of image-creation-25 have direct equivalent classes in French, so the only thing to do is to provide French examples with the right preposition(s), e.g. *with* in 25.3 has to be replaced in French with *de* or *avec*.

4.2.2 Principles on frames

So far, we have found two general differences between the coding of French and English frames in Verb∋net and VerbNet respectively.
The first one concerns "sub-frames", i.e. frames with missing complements such as *NP V* in 25.1 illustrated by *Smith was inscribing* which could be a sub-frame of e.g. *NP V NP.destination* (*Smith was inscribing the rings*). The coding of such sub-frames is dubious when based on introspection so it requires some corpus study. We don't know how this coding has been made in VerbNet and we don't have at our disposal enough French corpus data. So we decided for the time being to remove sub-frames from Verb∋net. For example in class remove-10.1, VerbNet encodes not only NP V NP PP.Source PP.Destination (*Doug removed the smudges from the tabletop*) but also NP V NP (*Doug removed the smudges*). Verb∋net only includes the first one; it is understood that the second one can be automatically inferred from the first one, without being (manually) validated[3].
The second one concerns the order of the complements. VerbNet sometimes encodes two frames which differ only by the order of the complements, e.g. in bring 11-3 the frames NP V NP PP.destination (*Nora brought the book to the meeting*) and NP V PP.destination NP (*Nora brought to lunch the book*). In French, the order of complements depends on a number of syntactic and semantic factors (Thuilier 2012), but it doesn't seem that it depends on a lexical factor, i.e. what is the lexical verb governing the complements. As a consequence, Verb∋net only records one frame in such cases, e.g. it only records *NP V NP PP.destination* (*Nora a apporté le livre au meeting*) with the direct object before the PP; it is understood that the other frame, *NP V PP.destination NP* (*Nora a apporté au meeting le livre*) can be automatically inferred from the first one.

4.2.3 Case by case work

In some cases, the second step in building Verb∋net is hard. There are two main reasons for that. First, there exist semantic differences which are taken into

[3]However, this principle concerning sub-frames is not applied for verbs which accept a single double-locative complement "from here to there (a single complement PP.source PP.destination)" without accepting a single source complement (PP.source), while accepting a single destination complement (PP.destination) : *Fred a transferré le vin de la cruche en pierre vers la cruche en terre cuite (Fred transferred the wine from the stone jar to the terra-cotta jar), *Fred a transferré le vin de la cruche en pierre (*Fred transferred the wine from the stone jar), Fred a transferré le vin vers la cruche en terre cuite (Fred transferred the wine to the terra-cotta jar.*

account in VerbNet but not in LVF or in LG. For example, among the verbs of Sending and Carrying (VerbNet super-class 11), the verbs in classes 11.3, 11.4 and 11.5 describe an accompanied motion (both the Agent and the Theme change location as in *Pamela drove packages to NY*), while those in classes 11.1 and 11.2 describe an unaccompanied motion (only the Theme changes location as in *Pamela sent packages to NY*). In the French resources, classes do exist for verbs with a change of location for a Theme caused by an Agent, but nothing is said about the Agent being or not being in motion. In the face of this difficulty, two solutions are possible: either make a study of French verbs of sending and carrying to distinguish accompanied and unaccompanied motions, or simply ignore this semantic difference. We opted for the second solution since this semantic difference does not appear to be relevant for a task such as semantic role labeling.[4] Ignoring this semantic difference leads us to adopt in Verb∋net a hierarchy for verbs of Sending and Carrying different from that in VerbNet: there is no equivalent in Verb∋net of class 11.4, the verbs belonging to this class being added to 11.1 or 11.2. Let us add that there is no French equivalent of class 11.3 made up of the two verbs *bring* and *take* with a deictically-specified direction (Levin 1993, page 135) since the French locative deictic *ici* and *là* don't work as *here* and *there*[5].

The second main source of difficulty comes from crucial differences between French and English. There exist translation problems between these two languages which are very well-known and documented, for example translation of motion verbs as illustrated in *John swam across the river* → *Jean a traversé la rivière à la nage* (lit. John crossed the river with the swim). We put aside those well-known cases here to discuss more subtle difficulties as illustrated with the verbs of change of possession. In VerbNet, there exist ten classes dedicated to these verbs. It seems that such a hierarchy cannot be adopted for French. Without going into all the details, let us underline the following points:

- The absence of dative and benefactive alternations in French means that the difference between VerbNet's classes 13.1 and 13.2 should probably not be kept.
- The semantic difference between 13.1 and 13.3 (namely HAS-POSSESSION versus FUTURE-POSSESSION) is perhaps too subtle and could be ignored.
- The preposition *with* in the frame corresponding to *Agent V Recipient with Theme* used in 13.4-1 and 13.4-2 has to be replaced with *en* and/or *de* according to the verb (e.g. *Luc livre Max en/*de lait*, *Luc équipe Max en/de téléviseurs*, *Luc dote Max *en/de téléviseurs*), which requires a reorganization into (sub-)classes.

All in all, it turns out that entering into the frame details has led us to revise the hierarchy of Verb∋net though we are trying to minimize the amount of revision in order to keep as much semantic information from VerbNet as possible.

5 Conclusion

We have presented a method for adapting the English syntactic and semantic resource VerbNet to a new language. This method combines the automation of structures transfer, automatic translation of the lexicon and linguistic expertise. We have applied this method to French and have reached a state where it is validated and the systematic work on each class is currently being realized. We are not able to give an evaluation of this resource since it is not yet completed. When it will be completed, we will make it freely available along with the web-based tool which makes it possible to collaboratively edit it.

In this work, we acknowledge the structural differences existing between languages: the class structure of Verb∋net does not follow exactly VerbNet. We keep track of such changes so that the differences between the two resources are explicit and well-documented. Thus they will be available for interacting with other resources through mappings, making our resource useful for multilingual applications.

This work is part of the ASFALDA[6] project which goals include the creation of a French FrameNet and mappings between it and other semantic resources, like LVF, LG and Verb∋net.

6 Acknowledgements

This work has been funded by the French national research agency (Agence Nationale de la Recherche, ANR) ASFALDA project under reference ANR-12-CORD-0023.

References

Bazrafshan, Marzieh and Daniel Gildea (2013). "Semantic Roles for String to Tree Machine Translation". In: *ACL 2013*.

Bonial, Claire, Orin Hargraves, and Martha Palmer (2013). "Expanding VerbNet with Sketch Engine". In: *Conference on Generative Approaches to the Lexicon (GL2013)*.

Boons, Jean Paul, Alain Guillet, and Christian Leclère (1976). *La structure des phrases simples en français : constructions intransitives*.

Dorr, Bonnie J, Mari Olsen, Nizar Habash, and Scott Thomas (2001). "LCS verb database". In: *Online Software Database of Lexical*.

Dubois, Jean and Françoise Dubois-Charlier (1997). *Les verbes français*. Larousse.

[4] Moreover, it seems that, for some English verbs, the Agent can be moving or not as reflected by the difference between VerbNet's classes 11.4 and 11.4-1.

[5] *Je suis là* (lit. *I am there*) can mean *Je suis ici* (lit *I am here*).

[6] https://sites.google.com/site/anrasfalda/

Exner, Peter and Pierre Nugues (2011). "Using semantic role labeling to extract events from Wikipedia". In: *DeRiVE 2011*.

Falk, Ingrid, Claire Gardent, and Jean-Charles Lamirel (2012). "Classifying French Verbs Using French and English Lexical Resources". In: *ACL 2012*.

Ferrer, Eva Esteve (2004). "Towards a Semantic Classification of Spanish Verbs Based on Subcategorisation Information". In: *ACL 2004: Student Research Workshop*. Barcelona, Spain.

Gale, William A., Kenneth W. Church, and David Yarowsky (1992). "Using bilingual materials to develop word sense disambiguation methods". In: *4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 101–112.

Gross, Maurice (1975). *Méthodes en syntaxe. Régime des constructions complétives*. Paris: Hermann.

Im Walde, Sabine Schulte (2006). "Experiments on the automatic induction of German semantic verb classes". In: *Computational Linguistics* 32.2, pp. 159–194.

Jentson, Indrek (2014). "VerbNet Workbench". In: *GWC 2014*.

Kipper-Schuler, Karin (2005). "VerbNet: A broad-coverage, comprehensive verb lexicon". PhD thesis. University of Pennsylvania.

Korhonen, Anna and Ted Briscoe (2004). "Extended lexical-semantic classification of English verbs". In: *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*.

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

Merlo, Paola, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria (2002). "A Multilingual Paradigm for Automatic Verb Classification". In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 207–214.

Messiant, Cédric, Kata Gábor, Thierry Poibeau, et al. (2010). "Acquisition de connaissances lexicales à partir de corpus: la sous-catégorisation verbale en français". In: *Traitement automatique des langues* 51.1, pp. 65–96.

Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda (2012). "An improved plagiarism detection scheme based on semantic role labeling". In: *Applied Soft Computing* 12.5, pp. 1493–1502. ISSN: 1568-4946.

Pradet, Quentin, Gaël de Chalendar, and Guilhem Pujol (2013). "Revisiting knowledge-based Semantic Role Labeling". In: *LTC'13*.

Saint-Dizier, Patrick (1996). "Constructing Verb Semantic Classes for French: Methods and Evaluation". In: *COLING 1996*.

Scarton, Carolina and Sandra Aluısio (2012). "Towards a cross-linguistic VerbNet-style lexicon for Brazilian Portuguese". In: *Workshop on Creating Cross-language Resources for Disconnected Languages and Styles Workshop Programme*, p. 11.

Shen, Dan and Mirella Lapata (2007). "Using Semantic Roles to Improve Question Answering". In: *EMNLP-CoNLL 2007*.

Sun, Lin, Anna Korhonen, Thierry Poibeau, and Cédric Messiant (2010). "Investigating the cross-linguistic potential of VerbNet: style classification". In: *COLING 2010*.

Surdeanu, Mihai, Sanda Harabagiu, Johns Williams, and Paul Aarseth (2003). "Using predicate-argument structures for information extraction". In: *Annual Meeting of the ACL 2003*, pp. 8–15.

Suzuki, Yoshimi and Fumiyo Fukumoto (2009). "Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering". In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*. Suntec, Singapore: Association for Computational Linguistics, pp. 32–40.

Swier, Robert and Suzanne Stevenson (2005). "Exploiting a Verb Lexicon in Automatic Semantic Role Labelling". In: *HLT-EMNLP 2005*.

Thuilier, Juliette (2012). "Contraintes préférentielles et ordre des mots en français". PhD thesis. Université Paris-Diderot.

Xie, Boyi, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer (2013). "Semantic Frames to Predict Stock Price Movement". In: *ACL 2013*.