Semi-automatic Propbanking for French

Claire Gardent and Christophe Cerisara

CNRS/LORIA, Nancy
Firstname.Lastname@loria.fr

Abstract

Although corpora annotated with both syntactic and semantic role annotations are now available for many languages, no such corpus is currently available for French. To address this shorcoming, we present a methodology for pre-annotating the semantic roles of verb arguments semi-automatically. We discuss the results obtained and give pointers for improving the approach.

1 Introduction

Corpora annotated with both syntactic and semantic role annotations permits training semantic role labellers i.e., systems which can identify and characterise (usually verbal) predicate/argument dependencies in text. For English, Propbank has been widely used [7] as well as Framenet [2]. As witnessed by the 2009 ConLL shared task "Syntactic and Semantic Dependencies in Multiple Languages", Propbank style corpora are available for many other languages such as in particular, German, Spanish, Catalan, Chinese, Korean. For French however, no such resource is currently available.

In this paper, we describe a methodology for pre-annotating verb arguments with semantic roles semi-automatically. Section 2 presents the methodology used, Section 3 discusses the results obtained and Section 4 concludes by summarising what remains to be done in order to obtain a fully annotated Propbank for French.

2 Methodology

We take as a starting point a corpus of newspaper articles annotated with dependency structures namely, the Paris 7 Dependency Treebank (P7Dep, [3]). This corpus gathers articles from Le Monde and contains 350 931 tokens, 12 351 sentences and 25 877 verb instances. The corpus was semi-automatically annotated with phrase structure trees [1] and manually verified, and the resulting treebank automatically converted to dependency structures [3]. The phrase structure annotations were furthermore exploited to automatically extract the TreeLex syntactic lexicon [6], which was then manually corrected.

To enrich the P7 dependency corpus with role labels, we first manually enrich the TreeLex lexicon with thematic grids so that each (verb, subcategorisation frame) pair is enriched with the appropriate syntax-to-semantics linking information (each syntactic argument is mapped to the appropriate semantic role). We then automatically label each verb instance with the subcategorisation frame used by that verb instance. Finally, we project from the enriched TreeLex lexicon, the thematic roles registered in this lexicon for that particular (verb, frame) pair. More specifically, we proceed in three steps as follows:

Adding thematic grids to Treelex. We use various existing resources (i.e., Dicovalence [8] and Propbank frame files [7]) to manually enrich the (verb,frame) pairs listed in Treelex with a linking between syntactic arguments and thematic roles. Since Treelex is a subcategorisation lexicon extracted from the P7 corpus, the verbs covered in this lexicon cover the verbs to be labelled in the corpus.

Associating P7 verb instances with subcategorisation frames. This is a preliminary step which permits projecting the thematic grid information contained in the enriched Treelex onto each verb instances in the P7 corpus. It consists in identifying the deep grammatical functions of each verb instance in the corpus. For instance, given the sentence *The cat is chased by the rat*, the surface agentive phrase *the rat* will be labelled as deep subject and the surface subject *the cat* as deep object.

Projecting Treelex thematic grids onto P7 verb instances. This steps builds on the previous two steps. For each verb instance in the P7 corpus, it projects the thematic grid information contained in the enriched Treelex onto the deep grammatical functions identified by the subcategorisation frame identification step. For instance, given the above sentence, it will project the a_0 label onto the deep subject *the rat* and the a_1 label onto the deep object *the cat*.

The procedure builds both on the parsed structure already present in the treebank and on the subcategorisation information present in Treelex which was extracted from this parsed corpus. The parse information facilitates the identification for each verb instance occurring in the corpus of its deep grammatical arguments. The subcategorisation information contained in Treelex once enriched with thematic grid permits an automated projection of thematic roles onto the parsed structure via the deep grammatical functions identified by the second step of the procedure. We now describe in more detail each of these steps.

2.1 Adding thematic grids to Treelex.

The aim of this first step is to associate each lexical entry (i.e., each (verb, subcategorisation frame) pair) in Treelex with a thematic grid and a mapping between grammatical functions and thematic roles. For instance, given the following lexical entry:

abîmer	SUJ:NP, OBJ:NP
(to damage)	*Ce champignon abîme les graines.*
	(This fungus damages the seeds.)

the aim is to produce the following enriched lexical entry:

abîmer	SUJ:NP:0 OBJ:NP:1
	Ce champignon abîme les graines.
	abîmer.01 damage.01 to harm or spoil
	0 agent, causer
	1 entity damaged

Resources used. To produce such entries, we use information from the P7 corpus, Dicovalence [8] and Propbank [7].

The *P7 corpus* gives us information about the usages of the verb in the form of sentences containing instances of it. We use this to help determine the meaning associated with each (verb, frame) pair.

Dicovalence is a subcategorisation lexicon which covers the most common French verbs and contains extensive information about each verb including in particular a translation to English. We use the Dicovalence translations of a verb as an indicator of its meaning and a bridge to the English Propbank.

Finally, the English *Propbank frames* associate a verb with a so-called roleset consisting of a verb meaning, a thematic grid and some illustrating examples. We use the Propbank frames to determine the thematic grid to be associated with a (verb,frame) pair in TreeLex given the verb meaning suggested by the English translation.

Manual editing. Given the information extracted from the P7 corpus (verb usage), Dicovalence (English translation for the verbs) and the Propbank frames (thematic grids), TreeLex is manually edited to associate each (verb,frame) pair with a meaning identifier, an english translation and an English gloss of that meaning, a thematic grid and a mapping between syntactic arguments and thematic role as illustrated by the enriched lexical entry for *abîmer* given above. The resulting files form the frame files of the French P7-Propbank.

This step of the procedure is time intensive with an average processing speed for a qualified linguist of 10 verbs per hour. Since there are 2 006 verbs in the Treelex lexicon, only a fraction of the verbs could so far be assigned a frame file thereby impacting semantic role labelling. We actually believe that a better way to proceed would be to first create verb classes and in a second step, to assign thematic grids to these classes rather than to isolated verbs. The automatic acquisition of verb classes from existing lexicons described in [5] is here particularly relevant. Indeed, we plan to apply this acquisition method to Treelex and to investigate in how far, the classes thus created group together verbs with identical thematic grids

and more particularly, identical mapping between syntactic arguments and thematic roles. In this way, instead of individually annotating 2 000 verbs, we would only need to annotate a few hundred classes.

2.2 Associating P7 verb instances with subcategorisation frames.

This step labels each verb argument with a deep grammatical function and a category consistent with the Treelex signature[1]. It then checks whether the resulting subcategorisation frame assigned to the verb is assigned to this verb by Treelex. Verbs labelled with a Treelex frame and verbs not labelled with a Treelex frame can then be distinguished and processed separately e.g., for debugging puroposes. More specifically, the frame labelling process proceeds in three steps namely, argument extraction and processing ; normalisation e.g. of passive and causative structures ; comparison with Treelex frames.

2.2.1 Argument extraction and processing

For each verb, a verb description is first produced which, based on the verb mood, on the verb auxiliary (if any) and on its arguments describes the verb environment (passive/active, infinitive/participial/finite form, causative embedding) and its arguments. For instance, given the P7 dependency annotations of the sentence shown at the top of Figure 2, the description associated with the verb *succèdera* (*to succeed*) will be as given in the lower part of the Figure. Additionally (though not shown by the graphical interface), the verb is marked as active.

This conversion from dependency annotations to verb description is implemented by a set of rewrite rules which assign each word related to the verb by an argumental relation, an argument description in the Treelex format i.e., a pair FUNCTION:CATEGORY where FUNCTION and CATEGORY are as listed in the Treelex part of Table 1. As indicated in this Table, the argumental relations taken into account to identify the arguments of a verb are the P7 relations suj, obj, de_obj, a_obj, p_obj, ats, ato and aff. For instance, the subject rule is as follows:

If *F = suj(V)* :

- If *cat(F)* $\in \{A, N, ET, CL, D, PRO, P + PRO, P + D\}$ then *SUJ:NP*
- If *cat(F)* = P then *SUJ:PP*
- If *cat(F)* = C then *SUJ:Ssub*
- If *cat(F)* = $VINF$ then *SUJ:VPinf*

Additionally, verb features are used to assign one or more of the following features to the verb description: infinitival, participial, passive and causative.

[1]The signature used to specify syntactic categories and functions in the P7 dependency treebank differs from that used in TreeLex. Hence the rules must map the P7Dep functions and categories to those used in TreeLex. This mapping is given in Table 1

TreeLex	description	**P7DEP**
SUJ	subject	suj
OBJ	object	obj
DE-OBJ	de-prepositional object	de_obj
A-OBJ	à-prepositional object	a_obj
P-OBJ	other prepositional object	p_obj
ATS	subject attribute	ats
ATO	object attribute	ato
refl	reflexive pronoun	aff
obj	affix	aff

TreeLex	**P7DEP**
NP	N
Ssub	C
PP	P
VPinf	VINF
il	il
en	en
CL	CL
AdP	ADV
y	y
VPpart	VPR
AP	A

Figure 1: Mapping P7/Treelex

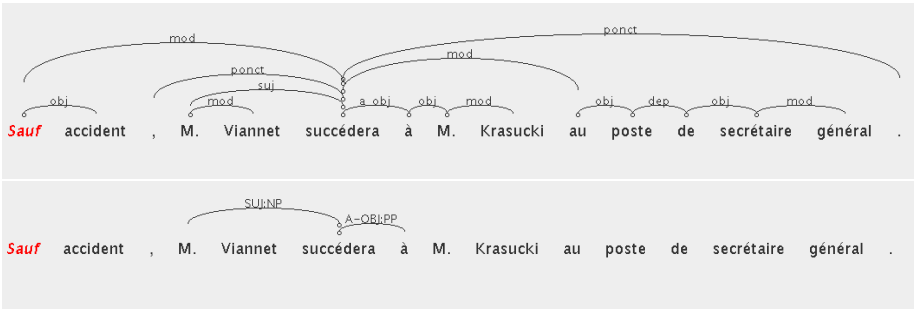

Figure 2: P7 dependency annotation and the resulting verb description for sentence *Barring accidents, M. Viannet will succeed M. Krasucki as general secretary.*

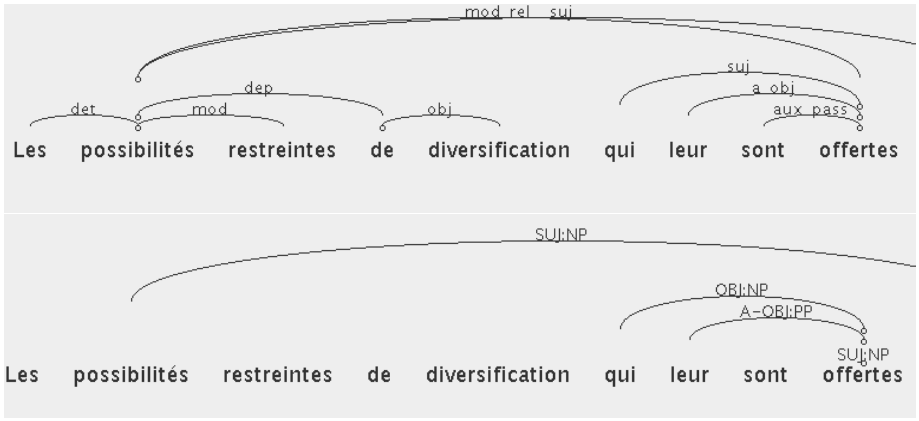

Figure 3: Normalising a passive in sentence *The limited diversification possibilities that are offered to them*

2.2.2 Normalisation

Given the verb description produced for each verb instance by the preceding step, the normalisation phase rewrites the frames of all verbs occurring in a passive, infinitival, participial or causative environment. The result is a frame assignment which relate each verb instance in the P7 dependency corpus to its arguments by an edge labeled with a deep grammatical function and a Treelex syntactic category. For instance, the frame assignment derived from the P7 dependency annotations for the verb *offertes* (*offered*) shown in the upper part of Figure 3 is as shown in the lower part of this figure. The surface subject *qui* (*that*) is labelled as an object NP, the dative clitic *leur* (*to them*) as a prepositional à -object and a subject NP is added.

2.2.3 Comparison with Treelex frames.

Finally, for each verb instance occuring in the P7 dependency corpus, the frame found by the above extraction procedure is checked against the frames associated with that verb by Treelex. If the frame exists in Treelex, the frame assignment is validated. Otherwise, the verb token is marked as having a non validated syntactic frame.

2.3 Projecting Treelex thematic grids onto P7 verb instances

The final step of the procedure assigns thematic roles to the deep arguments assigned to verb instances by the previous step using the Treelex lexicon enriched with thematic information described in section 2.1. For instance, given the en-

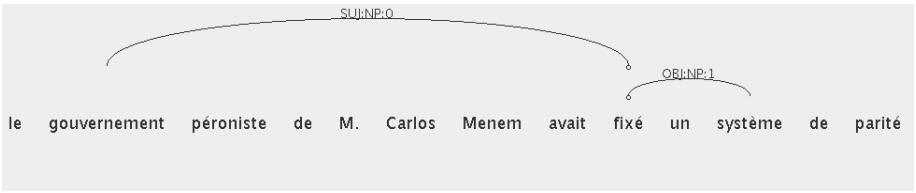

Figure 4: Final output

riched Treelex lexical entry for *fixer* shown below, the final output of our labelling procedure is as shown in Figure 4.

fixer SUJ:NP:0, (OBJ:NP:1)
fixer.01 establish, set
0 agent, setter
1 thing set
2 location, position, attribute
En mars , le gouvernement péroniste de M. Carlos Menem avait fixé un système de parité de 10000 australs pour 1 dollar, en vertu de la loi de convertibilité approuvée par le Congrès.
(*In march, the peronist government of M. Carlos Menem had set a parity system of 10000 australs for 1 dollar, under the convertibility low approved by the Congress.)*

3 Results and Evaluation

We applied the pre-annotation procedure described in the previous section to the P7 corpus annotated with dependency structures. This corpus contains 350 931 tokens, 12 351 sentences and 25 877 verb instances. 78% (25 113) of the verb instances were assigned a Treelex frame by the first step of the procedure and 42% (13815 tokens) could be labelled with semantic roles.

To analyse the output of each step of the role labelling procedure (frame extraction, frame validation by TreeLex, grid assignment), we developped some visualisation and annotation tools. We then carried out a pilote evaluation to assess precision (the correctness of results found) and recall (the proportion of correct results found).

3.1 Visualisation and annotation tools

To visualise and analyse the results of the semi-automatic annotation procedure described in the previous section, we developed a graphical interface which permits visualising intermediate and final results and separating sentences for which all verb tokens were successfully processed from sentences where at least one verb

token could not be processed [2]. More specifically, the menu provides 10 distinct views of the results, each view being named after (i) the annotations shown and (ii) the sentences they contain. The annotations can be one of the following. DEPS are the dependency annotations present in the initial P7 dependency corpus. In intermediate result files, dependency annotations are useful for checking whether a missing frame/grid stems from a parse error. RES are all the annotations produced by the annotation procedure described in the previous section. FRAMES are the subcategorisation frames extracted by the frame assignment procedure but not present in Treelex. This annotation level is useful for checking whether the frames found but not present in Treelex are either missing in Treelex or an incorrect result of the extraction procedure. TLFRAMES are the subcategorisation frames extracted by the frame assignment procedure and present in Treelex for the verb considered. These annotation level permits checking the precision of the extraction procedure (are the frames found and validated by Treelex actually the correct frames for the given verb tokens?). Finally, ROLES annotations are the thematic grids extracted by the SRL procedure. This annotation level when merged with the dependency annotations permits constructing the output Propbank.

Furthermore, the sentences contained in a file viewed can be any of the following. A P7 view will contain the entire P7 corpus; a NOTINTL view gathers sentences containing at least one verb whose extracted subcategorisation frame does not occur in Treelex. The ALLINTL views groups together sentences such that all verb tokens in those sentences were assigned a Treelex frame. A NOGRID view contain all the sentences where there is at least one verb for which no thematic grid could be extracted. Finally, the ALLINPBK view gathers sentences such that all verb tokens in those sentences were assigned a thematic grid.

3.2 Missing information (low recall)

There can be several reasons for the non identification of a frame or of a thematic grid.

A missing frame may stem from an incorrect dependency structure[3], a missing frame in Treelex or an incorrect/missing frame rewrite rule.

Missing thematic grids stem either from a missing frame (the verb token was not assigned a frame by the frame assignment procedure) or from a missing frame file (cf. section 2.1).

Decreasing the number of missing thematic grids requires improving the frame extraction step and extending the coverage of the frame files. As discussed in section 2.1, the latter is time intensive and will require a few more months for completion. Improving the former (the frame extraction step) requires analysing, quantifying and correcting the three possible sources of missing data (incorrect dependency

[2]This tool is available for download at `http://www.loria.fr/~cerisara/jsafran/index.html`

[3]This is turn may be due either to an incorrect annotation of the P7 treebank or to errors in the conversion script which project dependency structures from the initial constituency annotations.

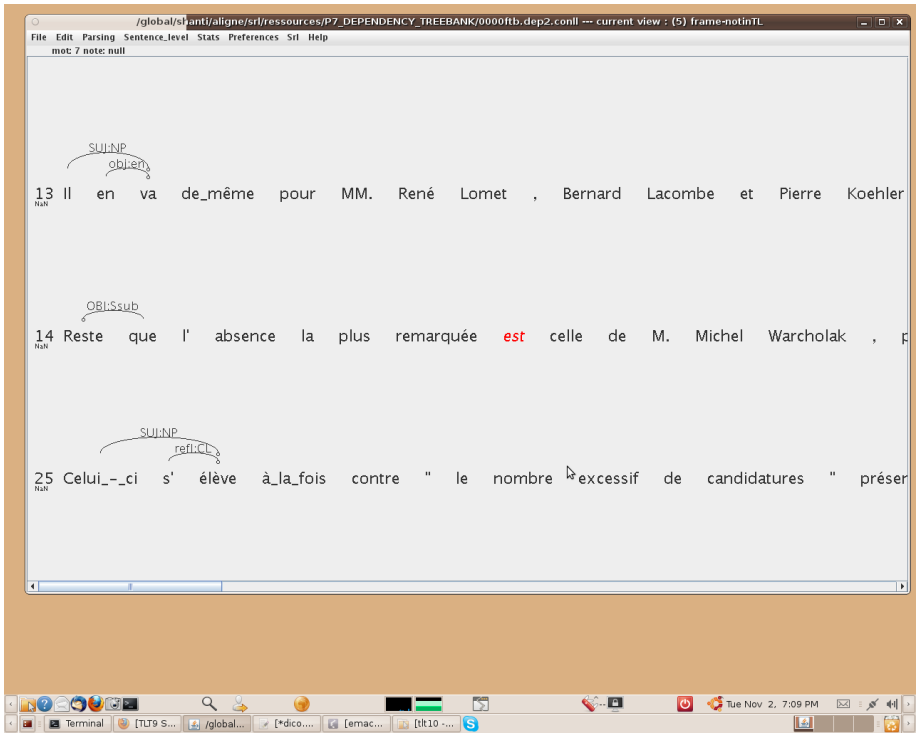

Figure 5: A FRAMES-NOTINTL view with the beginning of three sentences: (1) *It is the same for M. René Lornet, Bernard Lacombe and Pierre Koehler...* (2) *Still the most notable absence is that of M. Michel Wecholak, ...* (3) *He protested against both "the excessive number of candidates" ...* In the first and third case, the treebank fails to record a complement relation between the verb and a prepositional phrase (Pour- (*for*) and Contre-PP (*against*) respectively) so that the correct frame cannot be matched with any of the frames listed for “aller” (*to be*) and “s’élever” (*to protest*) respectively. In the second case, the frame found is probably correct but not listed in TreeLex.

structure, missing Treelex frame, incorrect/missing frame rewrite rule). To carry out such an investigation, we use the NOTINTL views. The FRAMES-NOTINTL view shows the frames found for those verb tokens for which the found frame is not in Treelex while the DEPS-NOTINTL view shows their dependency annotation. We use the second view (DEPS-NOTINTL) to identify incorrect frame assignment due to a parse error and the first (FRAMES-NOTINTL) to identify both errors in the extraction procedure and missing frames in Treelex. On a random sample of 50 verb tokens in the FRAMES-NOTINTL view (i.e., for which no TreeLex frame could be found), the results are as given in the following table.

Treebank error	22	44%
Missing Frame in Treelex	16	32%
Incorrect/missing frame rewrite rule	12	24%
TOTAL	50	100%

Manual inspection shows that treebank errors include erroneous dependency structures (often noun modifiers classified as de-objects or complements classified as modifiers) and incorrect lemmatisations (e.g., *secoué* (*shaked*) instead of *secouer* (*to shake*)). Missing Treelex frames often involves a mismatch between Treelex treatment of infinitival complements introduced by the preposition "de" and the treebank dependency structure annotation. Finally, incorrect/missing frame rewrite rules fall mainly into two cases namely, coordination and causative structures. We plan to extend the rewrite rules so as to correctly handle these structures too which should further increase the ratio of verb tokens for which a Treelex frame can be found. Provided Treelex and rewrite rule errors are fixed, the upper bound on the automatic identification of the subcategorisation frame of a verb token would thus approximates 88% the remaining errors being due to incorrect dependency annotations and missing information in TreeLex.

3.3 Erroneous frame assignment (precision)

Similarly, we analyse erroneous frame assignment by examining the ALLINTL views i.e., those sentences for which all verb tokens are assigned a frame validated by Treelex. On a sample of 50 verb tokens, 9 verb tokens were assigned an incorrect frame. Manual investigation showed the following distribution:

Treebank error	7	14%
Incorrect/missing frame rewrite rule	2	4%
Correct frame identification	41	82%

Again many of the treebank errors are noun complements categorised as verb de-objects.

3.4 Creating a training corpus for semantic role labelling

Our graphical interface also provides a functionality for merging dependency and thematic grid annotations so as to provide a training corpus for semantic role labelling. The format and content of this corpus is similar to the ConLL format [4].

4 Discussion, Conclusion and Perspectives

We have presented a semi-automated procedure for pre-annotating verb arguments with thematic roles in a dependency treebank for French. The automated part of the procedure (the identification of the deep grammatical functions of the verb arguments) accounts for 78% of the verb instances whereby the missing identifications are due mostly to errors in the dependency annotations (44% of the missing cases) and to missing information in the TreeLex lexicon (32%) of the missing cases). Only 12% of the missing identifications are due to errors in our procedure and these errors can relatively easily be fixed as the methods used (rewrite rules mapping surface to deep syntactic functions) are symbolic and the visualisation tools we developed, permit a detailed and systematic investigation of the error cases. Further, on the small sample we examined, precision (the proportion of correct mappings between surface and deep grammatical functions) reaches 82% with only 4% of the cases being due to errors in the annotation procedure, the remaining 14% being due to errors in the dependency annotations.

In sum, the automated part of our pre-annotation procedure displays a coverage and a precision which suggests that it can effectively support the development of a propositional bank for French. To ensure that the resulting annotated corpus supports the training of semantic role labellers, two points must be further pursued however.

First, Treelex must be fully augmented with thematic roles. As mentioned in section 2.1, this step could be enhanced by first producing a classification of French verbs which, as in the English VerbNet, groups together verbs, syntactic frames and thematic grids. [5] reports on an experiment in acquiring verb classes for French from existing lexical resources. This preliminary investigation suggests that Formal Concept Analaysis is an appropriate framework for bootstraping a verb classification for French from existing lexical resources and thereby to quickly associate thematic grids with sets of verb/frame pairs. In ongoing work, we are currently exploring how additionally taking into account syntactico-semantic features present in Dicovalence and in the LADL tables affects the classification and more specifically, whether such features permit creating verb classes that are sufficiently semantically homogeneous to contain mostly verbs that share the same thematic grid.

Second, adjuncts need to be dealt with. Indeed the present proposal focuses on so-called core arguments while Propbank style annotation requires that temporal, manner and locative adjuncts also be annotated. It remains to be seen in how

much the combination of adjunct rewrite rule with taxonomical knowledge about the semantic type of the arguments suffices to correctly label verb adjuncts.

References

[1] A. Abeillé, L. Clément, and A. Kinyon. Building a treebank for french. In *In Proceedings of the LREC 2000*, 2000.

[2] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics.

[3] M.-H. Candito, B. Crabbé, and M. Falco. Dépendances syntaxiques de surface pour le français. Technical report, Université de Paris 7, 2009.

[4] X. Carreras and L. Marquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the CoNLL-2005 Shared Task: Semantic Role Labeling*, pages 152–164, Ann Arbor, Michigan, June 2005.

[5] I. Falk and C. Gardent. Bootstrapping a classification of french verbs using formal concept analysis. In *Interdisciplinary workshop on verbs*, Pisa, Italy, 2010.

[6] A. Kupsc and A. Abeillé. Growing treelex. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 28–39. Springer, 2008.

[7] M. Palmer, P. Kingsbury, and D. Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[8] Karel van den Eynde and Piet Mertens. La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104, 2003.