D18 - Report on Propbank Experiment

Claire Gardent (CNRS/LORIA) et Mathieu Desnouveaux (INRIA Nancy)

Abstract.
Ce délivrable D18 présente une méthodologie pour enrichir le corpus en dépendances de Paris 7 avec des roles thématiques à la Propbank.
Mots clefs: grilles thématiques, étiquettage sémantique

Document Id	Passage/2010/D18/v1.0
Projet	ANR-06-MDCA-013 PASSAGE
Version	v1.0
Date	27 juillet 2010
État	final
Distribution	public

Consortium Passage

Ce document fait partie d'un projet de recherche financé par le programme MDCA de l'ANR sous la référence ANR-06-MDCA-013.

Institut National de Recherche en Informatique et en Automatique (INRIA)
Contact: Éric de la Clergerie
E-mail: Eric.de_la_clergerie@inria.fr

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)
Contact: Patrick Paroubek
E-mail: pap@limsi.fr

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
Contact: Claire Gardent
E-mail: gardent@loria.fr

CEA-LIST
Contact: Gaël de Chalendar
E-mail: Gael.de-Chalendar@cea.fr

Participants à ce rapport

Les partenaires suivants ont pris une part active au travail conduisant à l'élaboration de ce document, même si ils n'ont pas directement contribué à la rédaction de ce document:

LORIA
CEA-LIST
INRIA

Modifications

Version	Date	Auteur	Modifications
0.1	01.07.10	Claire Gardent	Création

Table des matières

Introduction

In the PASSAGE proposal, the WP6 work package aims to validate the lexical resources extracted from the merged results of the parsing campaign by integrating them into a parser and comparing the performance of that parser when used with its old lexicon on the one hand and with the newly acquired one, on the other.

As explained in Deliverable D10 however, the quality of the lexicon extracted from the parsed corpora remains fairly low in part because the merged results were unavailable at the time of the final extraction (June 2010). Because it was unclear when the merged results would be available, we opted instead for a Propbank experiment based on an existing, hand validated treebank namely, the Paris 7 treebank [CCF09]. In this report, we first show how the dependency version of that treebank can be enriched with semantic roles provided that the Treelex syntactic lexicon is extended to associate thematic roles with syntactic function. The method we present permits automatically annotating 74% of the verb instances with normalised syntactic frames. However, thematic roles could only be assigned to 40% of the verb tokens because the manual extension of the lexicon with thematic role was too time intensive to be completed. In the second chapter of the report, we therefore propose a method for creating verb classes for French that is promising in that (i) it has good distributional properties (it permits associating large sets of verbs with several frames at once) and good coverage (it covers most of the verbs and lexical entries in the Dicovalence syntactic lexicon). In ongoing work, we are exploring how additionally taking into account syntactico-semantic features present in two existing resources (namely, Dicovalence and 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. In this way, we plan to reduce the annotation load resulting from enriching a syntactic lexicon with thematic roles and to permit a faster, more consistent annotation of the P7 treebank with Propbank like semantic roles. The resources and tools developed by this workpackage are available at `http://talc.loria.fr/J-Safran-un-environnement-logiciel.html`.

Chapitre 1

Building a propbank for French : a pilot study

1.1 Introduction

Semantic role labelling (SRL, [GJ02]) consists in labelling the arguments of a predicate with thematic roles such *Agent, Patient, Instrument* or *Location*. Because it captures a core aspect of clausal meaning (namely, predicate/arguments structure), semantic role labelling is used in many applications that require broad coverage semantic processing such as information extraction, question answering and text summarisation.

Mostly, SRL systems are machine learning systems that learn from large amounts of data annotated with both syntactic and semantics (role) annotations. For English, Propbank has been widely used [PKG05] as well as Framenet [BFL98]. Corpora with thematic role annotations also exists for many other languages (e.g., German, Spanish, Catalan, Chinese, Korean). For French however, there is to date no such resource available. In this chapter, we describe a method for extending the P7 dependency treebank [CCF09] with Propbank style semantic roles. We start (Section 1.2) by presenting the methodology used. We then (Section 1.3) discuss the results obtained. We conclude (Section 1.4) by summarising what remains to be done in order to obtain a fully annotated Propbank for French.

1.2 Methodology

To enrich the P7 dependency corpus with role labels we proceed in three steps as follows :

Adding thematic grids to Treelex. We use Dicovalence and Propbank to associate each subcategorisation frame listed in Treelex with a thematic grid. Since Treelex is a subcategorisation lexicon extracted from the P7 corpus, this ensures an appropriate match between the verbs covered in the lexicon and 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 normalising the surface realisation variations

and 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 and for each verb instance in the P7 corpus, 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.

1.2.1 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
Ce champignon abîme les graines.

the aim is to produce the following enriched lexical entry :

abîmer SUJ :NP :A0 OBJ :NP :A1
Ce champignon abîme les graines.
abîmer.01 damage.01 to harm or spoil
0 agent, causer
1 entity damaged
2 instrument

Resources used. To produce such entries, we use information from the P7 corpus, Dicovalence [vdEM03] and Propbank [PKG05].

The *P7 corpus* gives us information about the uses of the verb in the form of sentences containing instances of it. In the above example for instance, it tells us that the verb *abîmer* occurs in the sentence

> A court d' argent - jugeant les prix du café trop bas - les planteurs n' ont plus protégé leurs arbres contre la rouille , ce champignon qui abîme les graines et jaunit les feuilles.

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. For

instance, Dicovalence associates the following meanings to the non pronominal form of the verb *abîmer*.

abîmer.1 damage, injure, harm, destroy
le sable risque d'abîmer votre appareil photo

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 . E.g.,

damage.01 to harm or spoil
*The events of April through June damaged the respect and confidence which most Americans previously had [*T*-1] for the leaders of China.*
Prices rose on the news that a sizable West German refinery was damaged [-1] in a fire , [*] tightening an already tight European market.*
0 agent, causer
1 entity damaged
2 instrument

Extracted Information. For each verb V in Treelex, we extract from these 3 resources :

- The P7 sentences containing an instance of V
- The Dicovalence translations and examples for each meaning of V it registers
- The thematic grid associated with the DV translations of V

We store this information in a file named V. For instance, the file *abîmer* contains the following information :

```
===abîmer (frames: 1; all verbs: 1)

TL Frames

   SUJ:NP, OBJ:NP (1)

P7 Sentences

   A_court_d' argent - jugeant les prix du café trop bas - les planteurs
n' ont plus protégé leurs arbres contre la rouille , ce champignon qui
abîme les graines et jaunit les feuilles .

DV Senses

   ============================
   abîmer.1 damage, injure, harm, destroy
   ============================
      DV example: le sable risque d'abîmer votre appareil photo

      PBK
      damage.01 to harm or spoil
          `` The events of April through June damaged the respect and
```

```
confidence which most Americans previously had [*T*-1] for the leaders
of China . ''
          Prices rose on the news that a sizable West German refinery
was damaged [*-1] in a fire , [*] tightening an already tight European
market .
          0 agent, causer
          1 entity damaged
          2 instrument


   ============================
   abîmer.2 decay, rot, get damaged
   ============================
   DV example: comme il a fait très chaud tous les fruits se sont abîmés


      PBK
      rot.01 to decompose or decay
          El Salvador is destroying more than 1.6 million pounds of food
that [*T*-1] had rotted in government warehouses , government officials
said [0] [*T*-2] .
          0 causer
          1 entity decaying
```

This text file is furthermore converted to an XML file respecting the DTD given in Annex **??**.

Manual editing. Finally, the verb files are manually edited to associate each Treelex frame with a meaning, 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 15 verbs per hour. Since there are 2 006 verbs in the Treelex lexicon, only a fraction of the verbs could be assigned a frame file thereby impacting semantic role labelling. Although we are currently continuing with manual frame file creation in order to improve coverage, 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 [GMdlC09] 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.

1.2.2 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. 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 la-

belled with a Treelex frame can then be distinguished and processed separately e.g., for debugging puroposes.

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.

Argument extraction and processing

For each verb, a description is produced based on the verb (mood, auxiliary) and on the arguments (grammatical function, part of speech, lemma) features. This description describes the verb environment (passive/active, infinitive/participial/finite form, causative embedding) and converts the argument description to the Treelex format using the mappings given in Figure 1.1. For instance, given the P7 dependency annotations of the sentence shown at the top of Figure 1.2, the description associated with the verb *succèdera* will be as given in the lower part of the Figure. Additionally (though not shown in the picture), the verb is marked as active.

The conversion from dependency annotation to verb description is implemented by a set of rewrite rules which assigns 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 Figure **??**. As indicated in this Figure, 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*cat1(F)* $\in \{A, N, ET, CL, D, PRO, P + PRO, P + D\}$ then *SUJ :NP*
- If *cat1(F)* = P then *SUJ :PP*
- If *cat1(F)* = C then *SUJ :Ssub*
- If *cat2(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.

The complete set of rewrite rules used to associate each verb instance in the P7 dependency corpus is listed in Appendix A.

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. As for the arguments, rewrite rules are used to convert the predicate/argument structures produced by the preceding step. 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* shown in the upper part of Figure 1.3 is as shown in the lower part of this figure. The surface subject *qui* is labelled as an object NP, the dative clitic *leur* as a prepositional à -object and a subject NP is added.

The set of rewrite rules used to normalise the frames is given in Appendix A.4.

TreeLex	P7DEP
SUJ	suj
OBJ	obj
DE-OBJ	de_obj
A-OBJ	a_obj
P-OBJ	p_obj
ATS	ats
ATO	ato
refl	aff
obj	aff

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

FIG. 1.1 – Mapping P7/Treelex

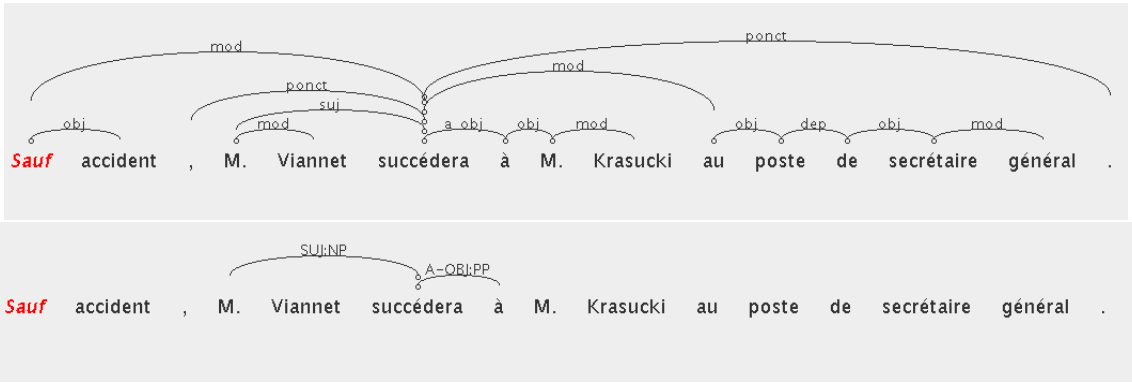

FIG. 1.2 – P7 dependency annotation and the resulting verb description

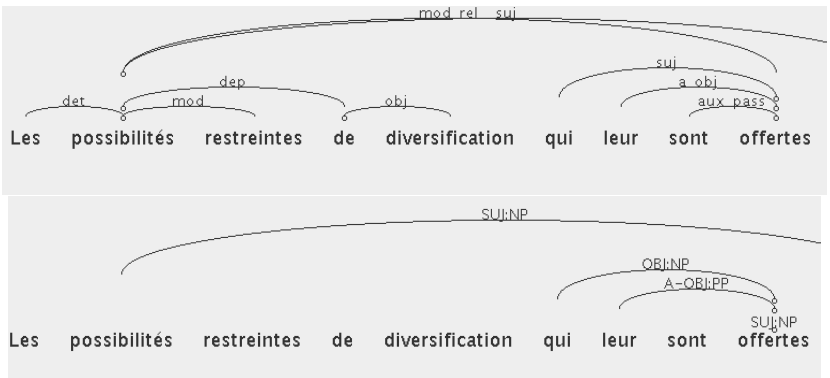

FIG. 1.3 – Normalising a passive

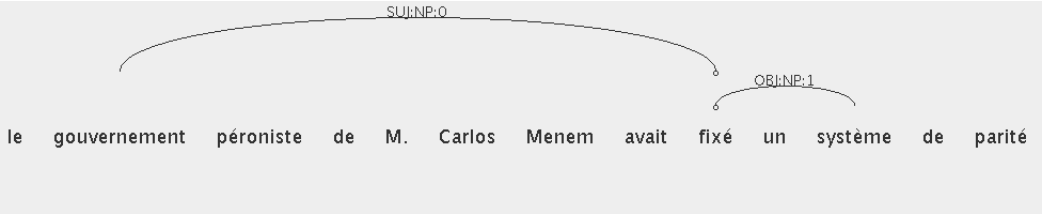

FIG. 1.4 – Final output

Comparison with Treelex frames.

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 frame assignment is checked manually and corrected if necessary.

1.2.3 Projecting Treelex thematic grids onto P7 verb instances

The final step 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 1.2.1. For instance, given the enriched Treelex lexical entry for *fixer* shown below, the final output of our labelling procedure is as shown in Figure 1.4.

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 .

1.3 Results and Evaluation

We applied the semantic role labelling 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 and carried out a pilote evaluation on 1 000 verb tokens aiming to assess precision (the percentage of incorrect results) and recall (the proportion of results not found).

1.3.1 Visualisation and annotation tools

J-SAFRAN (Java syntactico-semantic French analyzer, [CG09]) is a software environment which integrates the following functionalities :

- Syntactic analysis.

  Annotation. J-SAFRAN supports the manual annotation of text with a dependency graph

  Analysis. J-SAFRAN permits parsing French text with the MaltParser.

  Training. This functionality allows for training the MaltParser on a corpus annotated with dependency graphs.

  Evaluation. J-SAFRAN include functionalities which permit computing the ConLL evaluation metrics on the parser output (e.g., labelled attachment score or LAS)

- Semantic role labelling
  - Execution of the semantic role labelling described in this report
  - Visualisation of the intermediate/final results of the labelling procedure

J-SAFRAN can be downloaded from `http://www.loria.fr/~cerisara/jsafran/index.html`. It is implemented in Java. It is portable and easy to install under Windows, Linux or MacOsX. Annotation is WYSIWYG (What you see is what you get). The supported formats are text, XML (Syntex) and ConLL. For parsing, two models are provided with J-SAFRAN : one learned on the P7 dependency treebank with a LAS of roughly 88% and another learned on the ESTER corpus of broadcast news transcriptions with a LAS of 72%.

To visualise, annotate and analyse the results of our semantic role labelling procedure, we extend J-SAFRAN with a menu (called SRL) 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. More specifically, the SRL menu provides 10 distinct views named after the annotations shown and the sentences they contain. The annotation can be one of the following.

dep : 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 : all the annotations produced by the full SRL procedure.

frames : 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 : 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 ?)

srls : 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 :

P7 : the entire P7 corpus

notinTL : the sentences containing at least one verb whose extracted subcategorisation frame does not occur in Treelex

allinTL : sentences such that all verb tokens in those sentences were assigned a Treelex frame

nogrid : the sentences containing at least one verb for which no thematic grid could be extracted

allinPBK : sentences such that all verb tokens in those sentences were assigned a thematic grid

In total, the SRL menu allows the user to query for the following 10 files :

1. dep-P7 : all sentences in the P7 corpus. Annotation with dependency structures.
2. dep-notinTL : all sentences in the P7 corpus containing at least one verb token for which the extraction procedure could not identify a Treelex frame. Annotation in dependencies.
3. dep-allinTL : all sentences in the P7 corpus for which all verb tokens could be assigned a Treelex subcategorisation frame. Annotation in dependencies.
4. res-P7 : all sentences in the P7 corpus. Annotation with the results of the SRL procedure.
5. frame-notinTL : all sentences in the P7 corpus containing at least one verb token for which the extraction procedure could not identify a Treelex frame. Annotation with a subcategorisation frame.
6. TLframe-allinTL : all sentences in the P7 corpus for which all verb tokens could be assigned a Treelex frame. Annotation with Treelex subcategorisation frames.
7. dep-nogrid : all sentences in the P7 corpus containing at least one verb token for which the extraction procedure could not identify a thematic grid. Annotation in dependencies.
8. res-nogrid : all sentences in the P7 corpus containing at least one verb token for which the extraction procedure could not identify a thematic grid. Annotation with the result of the SRL procedure.
9. dep-allinPBK : all sentences in the P7 corpus for which all verb tokens could be assigned a thematic grid. Annotation in dependencies
10. srl-allinPBK : all sentences in the P7 corpus for which all verb tokens could be assigned a thematic grid. Annotation with the results of the SRL procedure.

1.3.2 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[1], 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 1.2.1).

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 1.2.1, the latter is

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

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 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 dep-notinTL view shows their dependency annotation. We use the second view (dep-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 sample of 50 verb tokens, 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%

Treebank errors include erroneous dependency structures (often noun modifiers classified as de-objects or complements classified as modifiers) and incorrect lemmatisations (e.g., *secoué* instead of *secouer*). 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 currently approximates 88%.

1.3.3 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 for the causes of these errors :

Treebank error	7	14%
Incorrect/missing frame rewrite rule	2	4%

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

1.3.4 Creating a training corpus for semantic role labelling

J-SAFRAN also provides a functionality ("merge" in the "srl" menu) 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 [CM05] and follows the following rules :

- Data files are UTF-8 encoded (Unicode).
- Data files contain sentences separated by a blank line.

- A sentence consists of one or more tokens, each one starting on a new line.
- A token consists of n fields described in the table below. Fields are separated by a single tab character. Space/blank characters are not allowed in within fields
- The fields used are :
  - ID, an integer identifying the token
  - FORM, the token form
  - CPOSTAG, a coarse POS tag as defined in the P7 dependency corpus cf. Appendix B
  - POSTAG, a detailed POS cf. Appendix B
  - FEAT, a list of morphological features
  - HEAD, a syntactic dependency
  - DEPREL, a dependency relation
  - FILLPRED, filled with "=Y" or left blank, this field indicates semantic predicates in this case, verbs
  - PRED, a sense identifier ; the PRED field is to be filled only for rows with FILLPRED= Y
  - as many APRED$_i$ fields/columns as there are verbs in the sentences PRED (where i is an integer)

1.4 Conclusion and Perspectives

The method and the tools described in the previous sections permit a fully automatic annotation of the P7 dependency treebank with Propbank style thematic roles. 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. The next chapter describes a first step towards this goal. The intuition underlying the proposed approach is that enriching lexical entries with thematic roles is best done at the verb classes level. Chapter 2 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 explore 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.

ID	Form	Lemma	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	_	_	FILLPRED	PRED		APRED
1	À	à	P	P	_	7	mod	_	_	_	_		
2	cette	ce	D	DET	g=fln=s|s=dem	3	det	_	_	_	_		_
3	époque	époque	N	NC	g=fln=s|s=c	1	obj	_	_	_	_		_
4	,	,	PONCT	PONCT	s=w	7	ponct	_	_	_	_		_
5	on	on	CL	CLS	g=mln=s|p=3|s=suj	7	suj	_	_	_	_		A0
6	avait	avoir	V	V	m=indln=s|p=3|t=impft	7	aux_tps	_	_	_	_		_
7	dénombré	dénombrer	V	VPP	g=mlm=partln=s|t=past	0	root	_	_	=Y	dénombrer.01		
8	cent_quarante	cent_quarante	D	DET	g=mln=pls=card	9	det	_	_	_	_		_
9	candidats	candidat	N	NC	g=mln=pls=c	7	obj	_	_	_	_		A1
10	.	.	PONCT	PONCT	s=s	7	ponct	_	_	_	_		_

TAB. 1.1 – Bla Bli.

Chapitre 2

Acquiring verb classes for French

2.1 Introduction

As discussed in Section 1.2.1, enriching Treelex with thematic grids is time intensive. In average, a qualified linguist can handle at most 15 verbs per hour. Moreover, assigning thematic grids to isolated verbs makes it difficult to ensure consistency across verbs. There is in particular, no easy way to ensure that the linguist assigns the same thematic grid to verbs that are syntactically and semantically similar. To remedy this shortcoming, we started investigating how to create verb classes that would gather together verbs sharing the same set of syntactic frames and ideally, the same thematic grid. More specifically, we aim to automatically create VerbNet like classes for French verbs on the basis of existing resources such as Dicovalence [vdEM03] and/or the LADL tables [Gro75a].

VerbNet ([Sch06]), is a large electronic verb classification for English which was created manually and classifies 3 626 verbs using 411 classes. Each VerbNet class includes among other things a set of verbs, a set of valency frames and a thematic grid. For instance, the *Hit-18.1* class associates verbs and frames as follows[1] :

[1]The Verbnet format for valency frames uses thematic roles rather than grammatical functions. We have used grammatical function here to preserve notation consistency and facilitate reading.

Class *Hit-18.1*	
Thematic roles	Agent [+int_control] Patient [+concrete] Instrument [+concrete]
Verbs	*batter, beat, bump, butt, drum, hammer, hit, jab, kick, knock, lash, pound, rap, slap, smack, smash, strike, tap*
Frames	SUJ :NP,P-OBJ :PP SUJ :NP,P-OBJ :PP,P-OBJ :PP SUJ :NP,OBJ :NP SUJ :NP,OBJ :NP,P-OBJ :PP SUJ :NP,DE-OBJ :Ssub

In this chapter, we present the result of a first experiment with FCA (Formal concept analysis, [GW98]) as a clustering method which we apply to Dicovalence. The resulting classification only groups together verbs with share a set of syntactic frames and does not yet consider thematic roles. We show however that FCA permits obtaining a verb classification that is promising in that it has good distributional properties (it permits associating large sets of verbs with several frames at once) and good coverage (it covers most of the verbs and lexical entries in the Dicovalence syntactic lexicon). In ongoing work, we are 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.

We start by outlining the intuition behind the proposal and describing the lexical resources used. We then show how FCA can be used to produce a verb classification and compare it with the English Verbnet.

2.2 Formal concept analysis

FCA is a classification technique which permits creating, from a so-called formal context, a concept lattice where concepts associate sets of objects with sets of attributes. Here, the concept objects will be verbs while the attributes will be syntactic frames and semantic features. Intuitively, a concept is a pair $\langle O, A\rangle$ such that all the objects in O have exactly the attributes in A and vice versa, all attributes in A are true of exactly all the objects in O. That is, our concepts will group together sets of verbs which share exactly the same set of syntactic and semantic features.

More formally, a formal context $\mathcal{K}$ is a triple $\langle \mathcal{O}, \mathcal{A}, R\rangle$ such that $\mathcal{O}$ is a set of objects, $\mathcal{A}$ a set of attributes and R a relation on $\mathcal{O} \times \mathcal{A}$. Given such a context, a concept is a pair $\langle O, A\rangle$ such that $O = \{o \in \mathcal{O} \mid \forall a \in A.\ (o, a) \in R\}$ and vice versa $A = \{a \in \mathcal{A} \mid \forall o \in O.\ (o, a) \in R\}$. A concept $C1 = \langle O1, A1\rangle$ is smaller than another concept $C2 = \langle O2, A2\rangle$ (written $C1 \leq C2$) iff $O1 \subseteq O2$ and $A1 \supseteq A2$. The set of all formal concepts of a context K together with the order relation $\leq$ form a complete lattice called the concept lattice of K. That is, for each subset

of concepts there is always a unique greatest common subconcept and a unique least common superconcept.

2.3 Lexical resources

2.3.1 Dicovalence

[vdEM03] is a syntactic lexicon for French verbs which lists among other things the valency frames of 3 936 French verbs. We use here a version of Dicovalence converted [Gar09] to the following format. Each verb is associated with one more valency frame which characterises the number and the type of the syntactic arguments expected by this verb. Further, each frame describes a set of syntactic arguments and each argument is caracterized by a grammatical function[2] and a syntactic category[3]. For instance, the frame of *Jean maintient ouvert le robinet / Jean maintain the tap open* will be SUJ :NP, OBJ :NP, ATO :XP .

2.3.2 The LADL tables

[Gro75b], [GL92] were specified manually over several years by a large team of expert linguists and contain syntactic and semantic information about French verbs. For instance, a table might state that the subject of all verbs in that table must be human ; or that the object is a destination, etc. The LADL tables group 5076 verbs into 61 distinct tables each table being associated with a defining valency frame and an informal description of the properties shared by verbs in that table[4].

2.3.3 VerbNet

[Sch06] is a verb classification for English which was created manually and classifies 3 626 verbs using 411 classes. Each VerbNet class includes among other things a set of verbs and a set of valency frames.

2.4 Acquiring verb classes

Our ultimate aim is to create a classification which facilitates the maintenance and verification of lexical verbal information such as in particular, valency frames and thematic grids. In the present paper however, we take an intermediate step towards that goal and seek to find a method for producing verb classifications which display the following properties.

[2]SUJ refers to the subject grammatical function, OBJ to the object, P-OBJ, A-OBJ and DE-OBJ describes prepositional objects introduced by any preposition, *à* ou *de* respectively and ATO indicates an object attribute.

[3]NP indicates a noun phrase, PP a prepositional phrase, CL a clitic and XP any major constituent

[4]The columns of the table give further more detailed information about each verb in the table but we do not use this information here.

Factorisation. The number of classes remains relatively small (no more than a few hundred) and in average, classes are balanced and well populated. That is, there are not too many classes with either very few frames or very few verbs.

Coverage. The classification covers most of the verbs and (verb, frame) pairs present in Dicovalence.

Similarity. The classes group together verbs sharing both a syntactic (frames) and a semantic (selectional restrictions, event type, argument structure) component

2.4.1 Creating the verb classification

The FCA lattice. To create verb classes which capture both a shared syntactic behavior (a shared set of valency frames) and a shared meaning component, we first build a concept lattice[5] based on the formal context $\langle V, F, R\rangle$ such that V is the set of verbs contained in the intersection of Dicovalence and the LADL tables, F is the union of the set of valency frames used in Dicovalence with the set of LADL table identifiers and R the mapping such that $(v, f) \in R$ if either Dicovalence or the LADL tables associates the verb v with the frame/table f.

Filtering. The resulting lattice contains 36065 concepts. To select from this lattice those concepts which are most likely to provide appropriate verb classes, we consider only concepts (i) whose attribute set contains at least one table identifier and one valency frame that is, which share both a syntactic and a semantic feature and (ii) that are intensionally stable ([Kuz07]). The intensional stability of a concept (V, F) is defined as $\sigma_i((V, F)) = \frac{|\{A \subseteq V | A' = F\}|}{2^{|V|}}$. Selecting concepts with high intensional stability yields classes which provide a good level of generalisation (their frame set is true of many verb sets).

2.4.2 Coverage.

One drawback with our filtering method is that since not all concepts are kept, some verbs and some frames might not be covered by the classification. In practice however, taking the 430 concepts with stability threshold 0.9995 (*Class430* in the following) and whose attribute set obey the set constraints (i.e., at least one table and one frame) yields a classification which covers 98.41% of the verbs, 25% of the frames and 83.17% of the (verb, frame) pairs. That is, the resulting classification covers most of the input data except for frames that have a rather low coverage due to many frames (in particular VPinf subject frames) with low frequency.

2.4.3 Comparison with Verbnet.

Table 2.1 gives a more detailed presentation of the impact of the stability threshold on the obtained classification. A threshold of 0.9995 yields a number of classes closest to that observed in Verbnet (430 against 411 in Verbnet). The main difference between Verbnet and our classification stems from the inventories of frames used. Although Dicovalence and Verbnet use approximately

[5] We used the Galicia Lattice Builder software (`http://www.iro.umontreal.ca/~galicia/`) to build the lattices

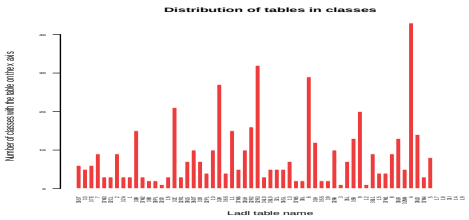

FIG. 2.1 – Distribution of tables in classes for *Class430*.

the same number of frames (116 and 117 respectively), many frames have a low frequency in Dicovalence so that our classification only retains 29 of the 116 initial Dicovalence frames. As a result, Verbnet has classes with a higher number of frames (average and maximum) and relatedly a lower number of verbs. Interestingly, finer grained classes are used in Verbnet where in particular, NP and PP categories are sometimes specialised with thematic roles (e.g., NP.patient vs NP.topic) and sentential arguments are differentiated into whether/how/what sentences. In future work, we intend to extend the classes and frames with thematic roles which might result in a classification distribution closer to that of Verbnet.

2.4.4 Factorisation.

Taking the 430 concepts with stability threshold 0.9995 (*Class430* in the following) and whose attribute set obey the set constraints (i.e., at least one table and one frame) yields a classification which covers 98.41% of the verbs, 25% of the frames and 83.17% of the (verb, frame) pairs. That is, it covers most of the input data except for frames that have a rather low coverage. Each class is associated with one or more semantic label (i.e., LADL table) and between 1 and 7 valency frames. Furthermore, the resulting classes each contain between 18 and 498 verbs. Overall thus, the classification obtained associates verb sets with an informative syntactico-semantic caracterisation ; groups together a satisfactory number of verbs and frames ; and permits covering a majority of verbs and (verb, frame) pairs present in Dicovalence.

We also plotted the LADL tables against the number of classes they include (Figure 2.1). For most tables (61%), less than 5 classes are identified. There are 5 tables which are assigned no class – these are all relatively small tables (around 20 verbs) for which no class could be found whose verbs were included in the set of verbs contained by the table.

2.4.5 Example class.

An example class extracted by this method associates the LADL tables 32RA (Make Adj_v), 8 (Verbs with sentential complement in *de*) and the frames SUJ :NP ; SUJ :NP,OBJ :NP ; SUJ :NP,DE-OBJ :Ssub with the verb set { blanchir (*to whiten*), bleuir (*to turn blue*), blêmir (*to turn pale*), pâlir (*to turn white*), rajeunir (*to become younger*), rosir (*to turn pink*), rougir (*to blush*), verdir (*to turn green*), vieillir (*to age*)}. That is, the class groups together verbs which indicate a change of state (mainly colour and age) and which can be used with and without object as well as with a sentential *de*-object.

Minimal stability	0.9999	0.9995	0.9990	VerbNet
Nb. of classes	340	430	500	411
Min. verbs	20	18	18	1
Max. verbs	498	498	498	383
Min. frames	1	1	1	1
Max. frames	5	7	7	25
Classes with 1 verb	0	0	1	29
Classes with 1 frame	41	45	49	44
Avg. class size (verbs)	78.5	70.13	66.16	14.96
Avg. class size (frames)	2.61	2.71	2.76	4.02
Avg. class size (harm. mean)	6.87	7.02	7.09	4.67
Verb coverage (%)	97.99	98.41	98.70	
Frame coverage (%)	17.74	18.28	18.28	
Verb-frame pairs coverage (%)	80.81	83.17	84.19	
Total number of verbs			3536	3626
Total number of frames			116	117

TAB. 2.1 – Some features of the verb classification depending on the chosen stability threshold.

2.5 Conclusion

Developing a verb classification by hand is time consuming and error prone. It also makes it difficult to ensure consistency within and across classes. The results presented in this paper suggest that FCA is an appropriate framework for bootstraping a verb classification for French from existing lexical resources. First, concepts naturally model the association between object (verbs) and attributes (syntactic and/or semantic features). Second, like fuzzy clustering, FCA permits "soft clustering" in that a data element may belong to several classes – a property of the produced classifications which is essential for our task since verbs are highly polysemic and may belong to several syntactic and/or semantic classes. Third, stable concepts and symbolic filtering on the attribute sets permit creating classes with good factorisation power (e.g., a few hundred syntactic classes to cover roughly 3 500 verbs) and linguistically sound, empirical content (good average number of verbs and frames within the classes).

Ongoing work concentrates on enriching the classification with additional features such as passivisation, reflexivisation, middle voice, etc. and evaluating the classes obtained in particular, wrt their ability to group together verbs with identical thematic grids.

Bibliographie

[BFL98] Collin F. Baker, Charles J. Fillmore, and John 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.

[CCF09] Marie-HÍène Candito, Benoit Crabbé, and Mathieu Falco. Dépendances syntaxiques de surface pour le français. Technical report, Université de Paris 7, 2009.

[CG09] C. Cerisara and C. Gardent. Analyse syntaxique du français parlé. In *Journée thématique ATALA Quels analyseurs syntaxiques pour le français ?*, 2009.

[CM05] 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.

[Gar09] Claire Gardent. Evaluating an Automatically Extracted Lexicon. In *4th Language & Technology Conference*, Poznan, Poland, 2009.

[GJ02] D. Gildea and D. Jurafsky. Automatic labelling of semantic roles. *Computational Linguistics*, 2002.

[GL92] A. Guillet and Ch. Leclère. *La structure des phrases simples en franÃ§ais. 2 : Constructions transitives locatives*. Droz, Geneva, 1992.

[GMdlC09] C. Gardent, C. Mouton, and E. de la Clergerie. D10 - technique d'acquisition de connaissances lexicales à partir de corpus analysés en syntaxe. Technical report, CNRS/LORIA, Projet ANR-06-MCDA-013 PASSAGE, 2009.

[Gro75a] M. Gross. *Méthodes en syntaxe*. Hermann, 1975.

[Gro75b] Maurice Gross. *MÃ©thodes en syntaxe*. Hermann, Paris, 1975.

[GW98] B. Ganter and R. Wille. *Formal Concept Analysis, Mathematical Foundations*. Springer-Verlag, Berlin, 1998.

[Kuz07] Sergei O. Kuznetsov. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4) :101–115, 2007.

[PKG05] M. Palmer, P. Kingsbury, and D. Gildea. The proposition bank : An annotated corpus of semantic roles. *Computational Linguistics*, 31(1) :71–106, 2005.

[Sch06] Karin Kipper Schuler. *VerbNet : A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2006.

[vdEM03] 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.

Appendices

Annexe A

Rewrite rules mapping P7 dependency structures to verb descriptions

A.1 P7/Treelex mapping

TreeLex	P7DEP
SUJ	suj
OBJ	obj
DE-OBJ	de_obj
A-OBJ	a_obj
P-OBJ	p_obj
ATS	ats
ATO	ato
refl	aff
obj	aff

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

A.2 Rewrite rules for the verb arguments

Le sujet Si *F = suj(V)* :

- Si *lemma(F) = falloir* alors *SUJ :il*
- Si *cat1(F)* $\in \{A, N, ET, CL, D, PRO, P + PRO, P + D\}$ alors *SUJ :NP*
- Si *cat1(F)* = P alors *SUJ :PP*
- Si *cat1(F)* $\in \{C, V\}$ alors *SUJ :Ssub*
- Si *cat2(F)* = $VINF$ alors *SUJ :VPinf*
- Sinon rien

L'objet Si F = obj(V) :

- Si *cat1(F)* $\in \{N, ET, CL, D, PRO, P, I\}$ alors *OBJ :NP*
- Si *cat1(F)* = A alors *OBJ :AP*
- Si *cat1(F)* = C alors *OBJ :Ssub*
- Si *cat1(F)* = $VPpart$ alors *OBJ :VPinf*
- Si *lemme(F)* $\in \{de, du\}$ et si (V2 = obj(F) et cat2(V2) = VINF) alors *OBJ :VPinf*
- Sinon rien

A-OBJ Si F = a_obj(V) :
- Si *lemme(F)* = *à* et si (*V2 = obj(F) et cat2(V2) = VINF*) alors *A-OBJ :VPinf*
- Si *lemme(F)* = *à* et si (*N = obj(F) et cat2(N) = N*) alors *A-OBJ :PP*
- Si *cat1(F)* $\in \{CL, P + PRO\}$ alors *A-OBJ :PP*
- Sinon rien

DE-OBJ Si *F = de_obj(V)* :
- Si *lemme(F)* $\in \{de, du\}$ et si (V2 = obj(F) et cat2(V2) = VINF) alors *DE-OBJ :VPinf*
- Si *lemme(F)* $\in \{de, du\}$ et si (N = obj(F) et cat2(N) = N) alors *DE-OBJ :PP*
- Si *cat1(F) = C* alors *DE-OBJ :Ssub*
- Si *lemma(F)* $\in \{dont, en, duquel, se\}$ alors *DE-OBJ :PP*
- Sinon rien

P-OBJ Si *F = p_obj(V)* :
- Si *lemme(F)* = P et si (V2 = obj(F) et cat2(V2) = VINF) alors *P-OBJ[P] :VPinf*
- Si *lemme(F)* = P et si (N = obj(F) et cat2(N) = N) alors *P-OBJ[P] :PP*
- Si *cat1(F) = C* alors *P-OBJ[P] :Ssub*
- Si *lemma(F)* = P alors *P-OBJ[P] :PP*
- Sinon rien

ATS Si *F = ats(V)* alors *ATS :XP*

ATO Si *F = ato(V)* alors *ATO :XP*

refl et obj Si *F = aff(V)* :
- Si *lemme(F) = en* alors *obj :en*
- Si *lemme(F)* = y alors *obj :y*
- Si *cat2(F) = CLR* alors *refl :CL*
- Si *cat2(F) = CL0* alors *OBJ :NP*
- Sinon rien

A.3 Rewrite rules operating on the verb features

Les traits et l'environnement du verbe sont extraits pour permettre la normalisation des structures passives et causatives.

Les traits et l'environnement du verbe sont extraits pour permettre la normalisation des structures passives et causatives.

- Si (*F = aux_pass(V)* et *cat2(V) = VINF*) alors *infinitive-passive(V)*
- Si *F = aux_pass(V)* alors *passive(V)*
- Si *F = aux_caus(V)* alors *causative(V)*
- Si *V = obj(V0)* et *lemma(VO)* $\in$ *Liste-vb-perception* alors *causatif(V)*
- Si *cat2(V) = VINF et lemma(V)* $\neq$ *faire* alors *infinitif(V)*
- Si *cat2(V) = VPR et lemma(V)* $\neq$ *faire* alors *participe-present(V)*
- Si *cat2(V) = VPP et V = mod(N) et lemma(V)* $\neq$ *faire* alors *participe(V)*

Liste-vb-perception : *voir, entendre, écouter* ..

A.4 Normalising frames

Given the arguments description produced by the preceding steps of the procedure, the frames of verbs occurring in a passive, infinitival, participial or causative context are normalised using the following set of rewrite rules.

A.4.1 Règle pour "Le monde daté 13 décembre 1999"

- Si *lemma(V) = dater* et *V = mod(*OBJ_1*)* et *DEOBJ*$_2$ *= mod(V)* alors *cadre(V) = SUBJ :NP* OBJ_1 *:NP DE-*OBJ_2 *:PP*

A.4.2 VP coordination

- Si *arguments(V) = ARGS* et *V = dep_coord(C)* et *C = coord(V0)*) et X_1 *= suj(V0)*) et *XP = cat(X)*) alors *cadre(V) =* SUJ_1 *:XP ARGS*

A.4.3 Infinitifs et participes

- Si *arguments(V) = ARGS* et *infinitif(V)* et *V = obj(V1)* et *cat(V1)* $\in \{V, VS\}$ et X^1 *= suj(V1)* et *cat(X) = XP*
  alors *cadre(V) =* SUJ^1 *:XP ARGS*
- Si *arguments(V) = ARGS* et *infinitif(V) ou participe-present(V)*
  alors *cadre(V) = SUJ :NP ARGS*
- Si *arguments(V) = ARGS (P-OBJ[par]*$_2$*) et V = mod(N)*$_1$ et *participe(V)* alors *cadre(V) =* SUJ_2 *:NP* OBJ_1 *:NP ARGS*

A.4.4 Passif

Si *infinitive-passive(V)* :

- Et Si *arguments(V) = ARGS (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :NP ARGS*

Si *passif(V)* :

- Et si *arguments(V) = (P-OBJ*^{1}*[par]) ARGS* et
  V = obj(V1) et *cat(V1)* $\in \{V, VS\}$ et X^2 *= suj(V1)* et *cat(X) = XP*
  alors *cadre(V) = SUJ*^{1} *:NP OBJ*^{2} *:XP ARGS*
- Et Si *arguments(V) = SUJ :NP (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :NP*
- Et Si *arguments(V) = SUJ :Ssub (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :Ssub*
- Et Si *arguments(V) = SUJ :VPinf (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :VPinf*
- Et Si *arguments(V) = SUJ :NP ARGS (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :NP ARGS*
- Et Si *arguments(V) = SUJ :Ssub ARGS (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :Ssub ARGS*
- Et Si *arguments(V) = SUJ :VPinf ARGS (P-OBJ[par])* alors *cadre(V) = SUJ :NP OBJ :VPinf ARGS*

A.4.5 Causatif

Si *causatif(V)* :

- Et Si *arguments(V) = OBJ :NP ARGS* et intransitif(V) alors *cadre(V) = SUJ :NP ARGS*
- Et Si *arguments(V) = OBJ :NP ARGS* et transitif(V) alors *cadre(V) = SUJ :NP OBJ :NP ARGS*

Annexe B

Tagets used by the P7 dependency corpus

Catégorie V	Catégorie V	Description / exemple verbe indicatif
VS	V	verbe subjonctif
VINF	V	verbe infinitif
VPP	V	participe passé
VPR	V	participe présent
VIMP	V	verbe impératif
NC	N	nom commun
NPP	N	nom propre
CS	C	conjonction de subordination
CC	C	conjonction de coordination
CLS	CL	clitique sujet
CLO	CL	clitique objet
CLR	CL	clitique réfléchi
P	P	préposition non amalgamée
P+D		préposition+dét : le lutin des alpages
P+PRO		préposition+prorel : le lieu auquel on pense
I	I	interjection
PONCT	PONCT	ponctuation
ET	ET	mots étrangers
ADJ	A	adjectifs non interrogatifs
ADJWH	A	adjectifs interrogatifs
ADV	ADV	adverbes non interrogatifs
ADVWH	ADV	adverbes interrogatifs
PRO	PRO	pronoms non interrogatifs ni relatifs
PROREL	PRO	pronoms relatifs
PROWH	PRO	pronoms interrogatifs
DET	D	déterminants non interrogatifs
DETWH	D	déterminants interrogatifs