Bootstrapping a Classification of French Verbs Using Formal Concept Analysis.

Ingrid Falk
INRIA/Nancy 2,
Nancy, France

Claire Gardent
CNRS/LORIA,
Nancy, France

Abstract

We use Formal Concept Analysis (FCA) to bootstrap a classification of French verbs. We show that the resulting classification has good factorisation power, compare it with the English Verbnet and report on a partial qualitative evaluation.

1 Introduction

Verb classifications have often been proposed which group together verbs with similar syntactic and/or semantic behaviour. On the practical side, verb classes permit capturing generalisations about verb behaviour thus reducing both the effort needed to construct a verb lexicon and the likelihood that errors are introduced when adding new entries. On the theoretical side, (Levin, 1993) has shown that syntax reflects semantics and consequently, that verbs that belong to a syntactic class can be shown to often share a semantic component.

In this paper, we explore the use of Formal Concept Analysis (FCA) to acquire classes for French verbs from available lexical resources. 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[1].

2 Formal concept analysis

FCA (Ganter and Wille, 1999) 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 $A = \{a \in \mathcal{A} \mid \forall o \in O.\ (o, a) \in R\}$. Two operators, both denoted by $'$, connect the power sets of objects $2^{\mathcal{O}}$ and attributes $2^{\mathcal{A}}$ as follows: $' : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{A}}, X' = \{a \in \mathcal{A} \mid \forall o \in X.\ (o, a) \in R\}$. The operator $'$ is dually defined on attributes. For a formal concept $\langle O, A\rangle \in O \times A$ we have $O' = A$ and $A' = O$. O is called the *extent* or *extension* and A the *intent* or *intension* of the formal concept.

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 $\mathcal{K}$ together with the order relation $\leq$ form a complete lattice called $\mathbb{K}$, the concept lattice of $\mathcal{K}$. That is, for each subset of concepts there is always a unique greatest common subconcept and a unique least common superconcept.

3 Lexical resources

We now present the linguistic resources used to build and evaluate a classification of French verbs namely, Dicovalence, the LADL tables and VerbNet.

Dicovalence (van den Eynde and Mertens, 2003) 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 (Gardent, 2009) as follows. Each verb is associated with one or more valency

[1] Other applications of FCA to linguistics and lexical resources are presented for eg. in (Priss, 2005) and (Valverde-Albacete, 2008).

frame characterising the number and type of the syntactic arguments expected by this verb. Further, each frame describes a set of syntactic arguments and each argument is characterised by a grammatical function[2] and a syntactic category[3]. For instance, the frame of *Jean maintient ouvert le robinet / Jean maintains the tap open* will be SUJ:NP, OBJ:NP, ATO:XP.

The LADL tables (Gross, 1975), (Guillet and Leclère, 1992) 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].

VerbNet (Schuler, 2006) 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.

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 aim at finding a method for producing verb classifications which display the following properties.

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

The FCA lattice. To create verb classes which capture both a shared syntactic behaviour (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 the set of objects V is the set of verbs contained in the intersection of Dicovalence and the LADL tables, the set of attributes F is the union of the set of valency frames used in Dicovalence with the set of LADL table identifiers and the relation 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. Not all these concepts are interesting verb classes however. In particular, many concepts only have 1 or 2 verbs and can hardly be viewed as classes. Similarly, concepts with few frames are less interesting especially if many of the verb subclasses of the extension of these concepts have more frames than there are in their intension. 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 (Kuznetsov, 2007). Intensional stability is a measure which helps discriminating potentially interesting patterns from irrelevant information in a concept lattice based on possibly noisy data. The intensional stability of a concept (V, F) is defined as $\sigma_i((V, F)) = \frac{|\{A \subseteq V | A' = F\}|}{2^{|V|}}$. In words, the stability of a concept (V, F) is defined as the number of those object subsets of V which have the same set of attributes as V divided by the total number of subsets of V. Intuitively, a more stable concept is less dependant on any individual object in its extent and is therefore more resistant to outliers or other noisy data items. 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]SUJ refers to the subject grammatical function, OBJ to the object, P-OBJ, A-OBJ and DE-OBJ describe prepositional objects introduced by any preposition, *à* or *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.

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

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.

5 Quantitative evaluation

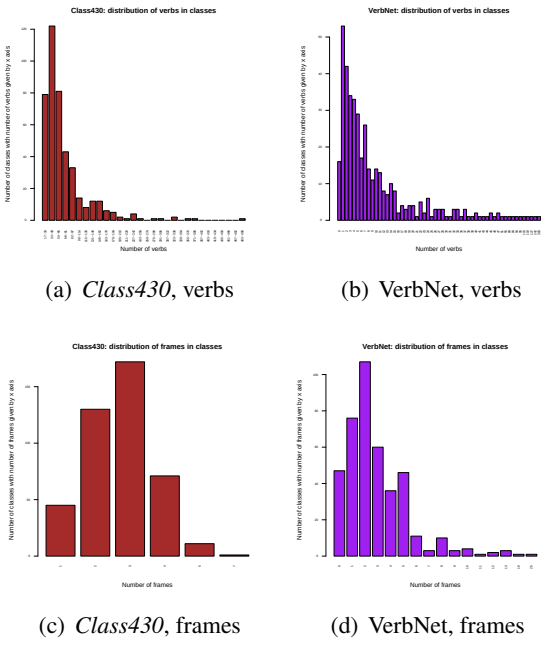
We first comment on the classification obtained based on a quantitative comparison with Verbnet.

5.1 Comparison with Verbnet.

Table 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). Fig. 1, a comparison of the distributions of verbs and frames in classes for VerbNet and *Class430*, shows that the distributions are similar, although VerbNet has more classes with a small number of verbs. The main difference between Verbnet and our classification stems from the inventories of frames used. Although Dicovalence and Verbnet use approximately 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.

5.2 Factorisation.

Each class is associated with one or more semantic label (i.e., LADL table) and between 1 and 7 va-



(a) *Class430*, verbs (b) VerbNet, verbs

(c) *Class430*, frames (d) VerbNet, frames

Figure 1: Number of classes with the number of verbs/frames given on the x-axis for VerbNet and *Class430*.

lency 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 characterisation; 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. 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.

5.3 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,DEOBJ:PP 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 senten-

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
Min. depth		2		1
Max. depth		6		4
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

Table 1: Some features of the verb classification depending on the chosen stability threshold.

tial *de*-object.

6 Qualitative evaluation

To explore the extent to which our classes group together verbs with identical thematic grids, we focus on psychological verbs i.e., verbs which, in the LADL tables, are described by table 4. Figure 2 shows the subgraph of our classification (*Class430*) rooted in the class with *table 4* as semantic feature. By inheritance all classes in this subgraph also have *table 4* as semantic feature.

Table 4 contains 616 verbs describing emotion or psychological verbs. All verbs in this table enter a transitive construction where the object is always human and the subject may be clausal (eg. *Que Luc agisse ainsi amuse Max /That Luc behaves this way amuses Max*). Because the subject of table 4 verbs may be phrasal, the EXPERIENCER is always the object (not the subject). Furthermore, the subject may accept both a (non-agentive) CAUSE and a (volitional) AGENT reading. Consequently, the thematic grid of verbs in this subclassification is [(CAUSE or AGENT), EXPERIENCER].

We now consider the subclassification in more detail and point out to several interesting ways in which the FCA approach interacts with polysemy and linking i.e., the mapping between syntactic arguments and thematic roles.

6.1 Polysemy

In Fig. 2, we outlined in blue the classes which have an additional table identifier in their attribute set and therefore may have an additional meaning[6]. For instance, class 4562 is associated not only with table 4 but also with table 32*C* which contains transitive verbs with a concrete object (Eg. *toucher le mur/touch the wall*). This suggests that the verbs in this class have both a psychological reading (Table 4, *toucher le public/move the audience*) and a concrete object reading (Table 32C, *toucher le mur/touch the wall*). More specifically, this suggest that verbs in class 4562 accept not one but 2 thematic grids and linkings namely:

Table 32C	Table 4
NP.AGENT NP.PATIENT	NP.CAUSE/AGT NP.EXP.
Jean touche le mur	*Jean touche le public*
Jean touches the wall	*Jean moves the audience*

6.2 Linking

Next we considered those classes (marked with a red font in Fig. 2) which are not characterised by an additional table and have at least 3 frames (marked with a red font in Fig. 2). That is, we consider classes which are semantically homogeneous (Table 4 only) and syntactically varied (several frames). For these classes, we examined whether it was possible to consistently determine linking i.e. to consistently assign thematic roles to syntactic arguments in the various frames.

This worked well for most of the 20 classes fulfilling our selection criterion (table 4, more than 3 frames). For instance, class 14650 (28 verbs) could be assigned the following linking information:

NP.CAUSE OR AGENT NP.EXPERIENCER
Jean irrite Marie/Jean irritates Mary
NP.CAUSE OR AGENT
Jean irrite/Jean irritates.
NP.EXPERIENCER, reflexive
Marie s'irrite/Marie irritates herself
NP.EXPERIENCER, PP.CAUSE OR AGENT, reflexive
Marie s'irrite contre Jean/Marie irritates herself against Jean.

Interestingly, Class 15856 departs from Class 14650 in that it groups together verbs for which the thematic role of the subject is ambiguous (Cause or Experiencer) when the verb is used in the intransitive form:

NP.CAUSE OR AGENT NP.EXPERIENCER
La douleur étouffe Marie/The pain suffocates Marie.
NP.CAUSE
La douleur étouffe/Pain suffocates.
NP.EXPERIENCER
Marie étouffe./ Marie suffocates
NP.EXPERIENCER, reflexive
Marie s'étouffe/Marie suffocates.

[6] In the LADL tables, the same verb occurring in different tables usually indicate that the verb has several possible meanings.

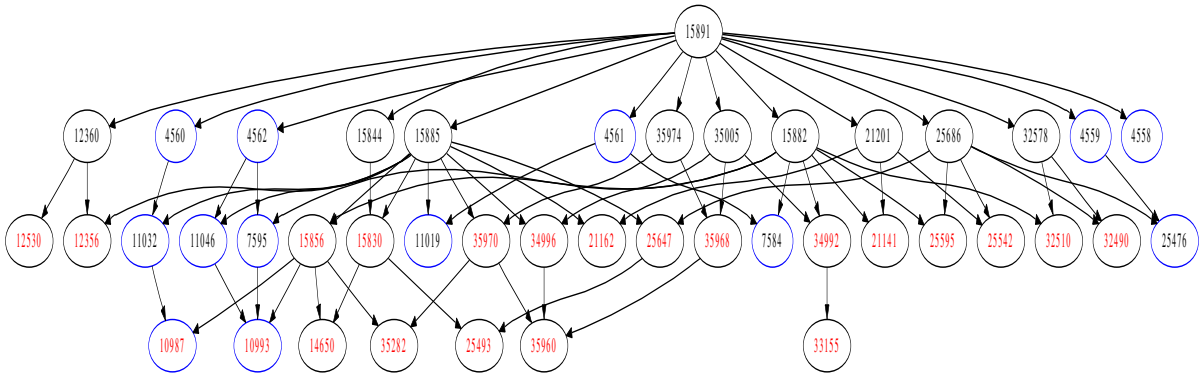


Figure 2: Hierarchic representation of verb/frame classes for LADL table 4

For the other eight classes with intransitive irreflexive frames this ambiguity did not appear, in particular the linking between syntactic argument and thematic role is straightforward in the three cases where the subject of the intransitive is clausal.

Finally, class 25647 (with 37 verbs) suggests that a more fine-grained representation of prepositional objects is needed to correctly determine linking. More specifically, information about preposition type (locative vs. beneficiary) is required to determine whether the EXPERIENCER role is realised by the object NP (*les jeunes*) or the prepositional object (*en moi*). Here, taking into account prepositions may help at separating the verbs of this class according to the syntactic realisation of EXPERIENCER.

NP.CAUSE NP.EXPERIENCER
Ceci exaspère/anime Marie.
NP.EXPERIENCER, reflexive
Marie s'exaspère/s'anime.
NP.CAUSE/AGENT, NP1, P-OBJ:PP
Elle exaspère [en moi]$_{\text{P-OBJ:EXPERIENCER}}$ *ce désir.*
Paul anime [les jeunes]$_{\text{NP1:EXPERIENCER}}$ *contre moi.*

To sum up, this case study shows that the proposed classification scheme permits associating thematic role with syntactic arguments for a large majority of classes.

7 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 polysemous 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). Fourth, a preliminary and partial qualitative evaluation suggests that the classes built adequately describe the association between verb sets, syntactic frames and thematic grids.

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

Acknowledgments

The research reported in this paper was partially supported by the French National Research Agency (ANR) in the context of the Passage project (ANR-06-MDCA-013). We would like to thank Yannick Toussaint for his suggestion to use stability as a filter as well as for general feedback and help on the topics addressed in this paper and Corinna Anderson for help and suggestions on several linguistic issues.

References

[Ganter and Wille1999] Bernhard Ganter and Rudolph Wille. 1999. *Formal concept analysis: Mathematical foundations*. Springer, Berlin-Heidelberg.

[Gardent2009] Claire Gardent. 2009. Evaluating an Automatically Extracted Lexicon. In *4th Language & Technology Conference*, Poznan, Poland.

[Gross1975] Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann, Paris.

[Guillet and Leclère1992] A. Guillet and Ch. Leclère. 1992. *La structure des phrases simples en français. 2 : Constructions transitives locatives*. Droz, Geneva.

[Kuznetsov2007] Sergei O. Kuznetsov. 2007. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):101–115.

[Levin1993] Beth Levin. 1993. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London.

[Priss2005] Uta Priss. 2005. Linguistic Applications of Formal Concept Analysis. In Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors, *Formal Concept Analysis*, volume 3626 of *Lecture Notes in Computer Science*, pages 149–160–160. Springer Berlin / Heidelberg.

[Schuler2006] Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

[Valverde-Albacete2008] Francisco J. Valverde-Albacete. 2008. Extracting Frame-Semantics Knowledge using Lattice Theory. *Journal of Logic and Computation*, 18(3):361–384, June.

[van den Eynde and Mertens2003] Karel van den Eynde and Piet Mertens. 2003. La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104.