Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Role Sets to French Verbs

Ingrid Falk[1], Claire Gardent[2]

[1] INRIA/Nancy Universités, Nancy (France)
[2] CNRS/LORIA, Nancy (France)

Abstract. We present an application of Formal Concept Analysis in the domain of Natural Language Processing: We give a general overview of the framework, describe its goals, the data it is based on, the way it works and we illustrate the kind of data we expect as a result. More specifically, we examine the ability of the stability, separation and probability indices to select the most relevant concepts with respect to our FCA application. We show that the sum of stability and separation gives results close to those obtained when using the entire lattice.

1 Introduction

Ideally natural language processing (NLP) applications need to analyse texts to answer the question of "Who did What to Whom". For computers to effectively extract this information from texts, it is essential that they be able to detect the events that are being described and the event participants. Because events are mostly lexicalised using verbs, one ingredient that is essential for such systems is detailed knowledge about their syntactic and semantic behaviour. It has been shown (Briscoe and Carroll (1993), Carroll and Fang (2004)) that detailed subcategorisation information (that is, information about the number and the syntactic type of verb complements) is crucial in enhancing their linguistic coverage and theoretical accuracy. However this syntactic information is not sufficient to specify "Who did what to Whom" because it does not allow to identify the thematic roles participating in the event described by the verb. For example in *John threw a ball to Mary* the syntactic analysis of the sentence would not allow to identify John which is the syntactic subject of the sentence as the Agent or Causer of the *throwing* event, Mary, syntactically the prepositional object as the Destination and *ball* (the object) as the item being *thrown*.

To help computer systems in this task of understanding and representing the full meaning of a text, verb classifications have been proposed which group together verbs with similar syntactic and semantic behaviour, ie. which associate groups of verbs with subcategorisation frames showing the syntactic constructions the verbs may appear in and sets of thematic roles which represent the participants in an event described by the verbs in the group.

ISBN 978–2–905267–78–8, INRIA Nancy – Grand Est and LORIA, 2011.

For English, there exist several large scale resources providing verb classes (*eg.* Framenet Baker et al. (1998) and VerbNet Schuler (2006), the classification we use in our framework) in a format that is amenable for use by natural language processing systems. For example for the verb *throw* the corresponding VerbNet class shows that the participants in a *throwing* event are an Agent, a Theme (the thing being *thrown*), a Source and a Destination. In addition, the VerbNet class provides the syntactic constructions the verb can occur in (eg. SUBJECT(*John*) V(*throws*) OBJECT(*a ball*) PREPOBJECT(*to Mary*)) and shows how the participant roles can be realised as syntactic arguments: In the example above the Agent (*John*) is realised syntactically as SUBJECT, the Theme (*the ball*) as OBJECT and the Destination (*to Mary*) as prepositional object (PREPOBJECT).

For French however, existing verb classes are either too restricted in scope (Volem Saint-Dizier (1999)) or not sufficiently structured (the LADL tables Gross (1975)) to be directly useful for NLP. Even though recently other large coverage syntactic-semantic resources for French have been made available (Tolone (2011) as well as further processed versions of Dubois and Dubois-Charlier (1997), Hadouche and Lapalme (2010)) the terminology and linguistic formalisms they are based on is often still hardly compatible with the methods and tools currently used in the NLP community.

In this paper we present a method for providing a VerbNet style classification of French verbs which associates verbs with syntactic constructions on the one hand and sets of semantic role sets (the set of semantic roles participating in the event described by the verb) on the other. To obtain this classification, we build and combine two independent classifications. The first is semantic and is obtained from the English VerbNet (VN) by translation, the second is syntactic and is obtained by building an FCA (Formal Concept Analysis) lattice from three, manually validated syntactic lexicons for French. The first associates groups of French verbs with the semantic roles of the English VN class. The second associates groups of French verbs (the concept extent) with syntactic constructions (concept intent). We then merge both classifications by associating with each translated VN class, the FCA concept whose verb set yields the best F-measure with respect to the verb sets contained in each translated VN class. We thus effectively associate the set of semantic roles of the VN class to the group of French verbs and the syntactic information given by the FCA concept.

In the past several linguistic FCA applications have been presented, as Priss (2005) shows in her overview. For example, Sporleder (2002) describes an FCA based approach to build structured class hierarchies starting from unstructured lexicon entries while the features used for building classes in the approach presented in (Cimiano et al., 2003) are collected from a corpus. Our approach (based on earlier work presented in Falk et al. (2010), Falk and Gardent (2010)) is concerned with building a lexical resource based on lexicons and is therefore related to the FCA approach in (Sporleder, 2002). However, the features we use are different. In addition we explore the use of concept selection indices to filter the concept lattices and finally relate the formal concepts we obtain to other classes

obtained by a clustering approach based on different numeric features extracted from lexicons and English-French dictionaries.

In the following we first introduce the terminology and data used in our application domain. Next we describe how we associate groups of French verbs with syntactic information using Formal Concept Analysis (Section 3). As the resulting concept lattice has a very large number of concepts which are mostly not useful verb classes we explore methods to select the concepts most relevant to our application (Section 4). We show in particular that selecting only $\sim 10\%$ of the concepts of the lattice using indices proposed in Klimushkin et al. (2010) gives results close to those obtained when using the entire lattice. We then show how we build the translated VerbNet classes and how they are mapped to the previously pre-selected FCA concepts (Section 5). Finally in Section 6 we present the kind of associations we obtain by our method.

2 Linguistic Concepts and Resources

Our aim is to build a lexicon associating groups of French verbs with:

1) the syntactic constructions the verbs of this group may appear in,
2) the semantic roles participating in an event described by a verb of this group.

Syntactic constructions a verb may occur in are described using *subcategorisation frames* (SCF) and are usually part of a lexical entry describing the verb. A subcategorisation frame (SCF) characterises the number and the type of the syntactic arguments expected by a verb. Each frame describes a set of syntactic arguments and each argument is characterised by a grammatical function (*eg.* SUJ - subject, OBJ - direct object etc.) and a syntactic category (NP indicates a noun phrase, PP a prepositional phrase, etc.). For example *John throws a ball to Mary.* is a possible realisation of the subcategorisation frame SUJ:NP V OBJ:NP POBJ:PP.

The semantic (thematic) roles are the participants in an event described by a particular verb. To date there is no consensus about a set of semantic roles or a set of tests determining them. There may be a general agreement on a set of Semantic Roles (eg. Agent, Patient, Theme, Instrument, Location, etc.) but there is substantial disagreement on when and where they can be assigned (Palmer et al., 2010). Thus each of the well known resources (FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), VerbNet (Schuler, 2006), LVF (Dubois and Dubois-Charlier, 1997)) providing semantic role information have their own semantic role inventory. In our work we chose the VerbNet semantic role inventory for several reasons:

1. VN semantic roles provide a compromise between generalisation and specificity in that they are common across all verbs[3] but are still able to capture specificities of particular classes.

[3] in contrast to FrameNet Baker et al. (1998) and PropBankPalmer et al. (2005) roles.

2. VN roles are among those generally agreed upon in the community.
3. None of the other resources provide the link between syntactic arguments and semantic roles across different verbs.
4. Semantic roles are expected to be valid across languages and by using the same role inventory as for English we hope to leverage some of the substantial research done for English and link syntactic information for French with semantic information provided by the English classes. Our method allows us to detect groups of French verbs with the same role set as some English VerbNet class and gives information about how these semantic roles are realised syntactically in French.

Figure 1 shows an excerpt of the *throw-17.1* VerbNet class, with its verbs, thematic roles and subcategorisation frames.

verbs (32): kick, launch, throw, tip, toss, ...
sem. roles: AGENT, THEME, SOURCE, DESTINATION

frames (8):

SCFs	**sem. roles**
Subject V Object	Agent V Theme
John throws a ball	
Subject V Object PrepObject	Agent V Theme Destination
John throws a ball to Mary	
Subject V Object Object	Agent V Destination Theme
John throws Mary a ball	
etc.	

Fig. 1: Simplified VerbNet class *throw-17.1*.

Thus, from this data an English NLP system analysing the sentence *John threw a ball to Mary* could infer the semantic roles involved in the event, namely those given by the VerbNet class. It could also detect the possible semantic roles realised by the syntactic arguments: It would know that the subject is a realisation of the Agent semantic role, the object of the Theme or Destination semantic roles, etc.

3 Associating French Verbs with Subcategorisation Frames

To associate French verbs with syntactic frames, we use the FCA classification approach where the objects are verbs and the attributes are the subcategorisation frames associated with these verbs by the subcategorisation lexicon to be described below.

3.1 Subcategorisation Lexicons

Subcategorisation information is retrieved from three existing lexicons for French: *Dicovalence* van den Eynde and Mertens (2003), *the LADL tables* Gross (1975),

Guillet and Leclère (1992) and finally *TreeLex* Kupść and Abeillé (2008). Each of these was constructed manually or with an important manual validation by linguists. The combined lexicon covers 5918 verbs, 345 SCFs and has a total of 20443 ⟨verb, frame⟩ pairs. Table 1 shows sample entries in this lexicon for the verb *expédier (send)*. Using the Galicia Lattice Builder software[4], we first build

Verb: *expédier*	
SCF	**Source info**
SUJ:NP,DUMMY:REFL	DV:41640,41650
SUJ:NP,OBJ:NP	DV:41640,41650;TL
SUJ:NP,OBJ:NP,AOBJ:PP	TL
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP	LA:38L

Table 1: Sample entries in subcategorisation lexicon for verb *expédier (send)*.

a concept lattice based on the formal context $\langle V, F, R\rangle$ such that:

- V is the set of verbs in our subcategorisation lexicon. We ignore verbs with only one SCF as they will result in classes associating verbs with a unique frame.
- F is the set of subcategorisation frames (SCFs) present in the subcategorisation lexicon,
- R is the mapping such that $(v, f) \in R$ iff the subcategorisation lexicon associates the verb v with the SCF f.

The resulting formal context is made of 2091 objects (verbs) and 238 attributes (frames), giving rise to a lattice of 12802 concepts. Clearly however not all these concepts are interesting verb classes. Classes aim to factorise information and express generalisations about verbs. Hence, concepts with few (1 or 2) verbs can hardly be viewed as classes and similarly, concepts with few frames are less interesting.

To select from this lattice those concepts which are most likely to provide the most relevant verb-frame associations, we explore the use of three indices for concept selection: *concept stability*, *separation* and *probability* which have been proposed and analysed in (Klimushkin et al., 2010). In Section 4.2 we investigate which of these indices performs best in the context of our application. We then use the best performing concept filtering method to select the most relevant concepts with respect to our data. For each translated VN class we then identify among the selected FCA concepts the one(s) with best f-measure between precision and recall. For a translated VN class C_{VN} (consisting of French verbs) and the extent (verb set) of an FCA concept C_{FCA} precision, recall and f-measure are computed as follows: $R = \frac{|C_{VN} \cap C_{FCA}|}{|C_{VN}|}, P = \frac{|C_{VN} \cap C_{FCA}|}{|C_{FCA}|}, F = \frac{2RP}{R+P}$
The translated VN class is then associated with the FCA concept(s) with best F-measure. Thus the verbs in the FCA concept are effectively associated with the thematic roles of the translated class and at the same time with the syntactic subcategorisation frames in the intent (attribute set) of the FCA concept.

[4] `http://www.iro.umontreal.ca/~galicia/`

4 Filtering Concept Lattices

The lattices we have to deal with are very large and many of the concepts do not represent valid verb classes. To select those concepts which are most relevant in the context of our application the concept lattice needs to be filtered. Klimushkin et al. (2010) propose three indices for selecting relevant concepts in concept lattices built from noisy data: *concept stability, separation* and *probability*. In this section, we investigate which of these indices works best for our data.

Concept stability is a measure which helps discriminating potentially interesting patterns from irrelevant information in a concept lattice based on possibly noisy data. The stability of a concept $C = (V, F)$ is the proportion of subsets of the extent V which have the same attribute set F as V:

$$\sigma((V,F)) = \frac{|\{A \subseteq V \mid A' = F\}|}{2^{|V|}}.^5 \tag{1}$$

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.

Concept separation indicates the significance of the difference between the objects covered by a given concept from other objects and, simultaneously, between its attributes and other attributes:

$$\mathfrak{s}((V,F)) = \frac{|V|\,|F|}{\sum_{v\in V}|\{v\}'| + \sum_{f\in F}|\{f\}'| - |V|\,|F|}. \tag{2}$$

Intuitively we expect a concept with high separation index to better sort out the verbs it covers from other verbs and simultaneously the frames it covers from other frames. Whereas concept stability is a measure concerned with either objects or attributes, separation gives information about objects and attributes at the same time.

Concept probability. For an attribute $a \in A$, the attribute set, we denote by p_a the probability of an object to have the attribute a. In practise it is the proportion of objects having a: $p_a = \frac{|\{a\}'|}{|O|}$, where O denotes the set of objects.

For $B \subseteq A$, we define p_B as the probability of an arbitrary object having all attributes from B: $p_B = \prod_{a\in B} p_a$. This formulation assumes the mutual independence of attributes. Based on this, and denoting $n = |O|$ we obtain the following formula for the probability of B being closed:

$$p(B = B'') = \sum_{k=0}^{n} p(|B'| = k, B = B'') \tag{3}$$

$$= \sum_{k=0}^{n} \left[\binom{n}{k} p_B^k (1-p_B)^{n-k} \prod_{a\notin B} (1-p_a^k) \right] \tag{4}$$

[5] Here and in the following $'$ represents the operator on the power sets of objects: $' : 2^{\mathcal{O}} \to 2^{\mathcal{A}}, X' = \{a \in \mathcal{A} \mid \forall o \in X.\ (o,a) \in R\}$ and dually on that of attributes.

A small $p(B = B'')$ suggests a small probability of the attribute combination B to be a concept intent by chance only (and $p(B = B'') \approx 1$ that there is a high probability that the combination is a concept intent by chance). However, this reasoning is based on the independence of the attributes, which in our particular case can not be warranted.

4.1 Computing Stability, Separation and Probability Indices.

Stability. Calculating stability is known to be NP-complete (Kuznetsov, 2007), however Jay et al. (2008) show that when the concept lattice is known it can be computed efficiently by a bottom-up traversal algorithm introduced in (Roth et al., 2006). This is the algorithm we used to compute concept stability.
Separation can be computed in $\mathcal{O}(|O|+|A|)$ time, where O and A are the object and attribute sets respectively. Computing separation is the least prohibitive of the three indices.
Probability. Klimushkin et al. (2010) show that computing probability of only one concept involves $\mathcal{O}(|O|^2 \cdot |A|)$ multiplication operations which is computationally very costly. With the computational means at our disposal it was not possible for us to compute the concept probabilities. We therefore computed approximations derived as follows:
First, we consider $\prod_{a \in B}(1 - p_a^k) \approx 1$ for $k > 40$. In view of this, Equation (4) becomes:

$$p(B = B'') = \sum_{k=0}^{40} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{a \notin B} (1 - p_a^k) \right] \tag{5}$$

$$+ \sum_{k=41}^{n} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \right] \tag{6}$$

As $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$, Term (6) can be rewritten as:

$$1 - \sum_{k=0}^{40} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \right] = \tag{7}$$

$$1 - F(40; n, p_B). \tag{8}$$

$F(k; n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-1}$ is the cumulative distribution function of the binomial distribution[6] and can be computed using various statistical software packages. Term (5) can also be computed more easily considering that $\binom{n}{k} p_B^k (1 - p_B)^{n-k}$ are binomial densities the computation of which is also provided by statistics software[7]

[6] Source Wikipedia: `http://en.wikipedia.org/wiki/Binomial_distribution`
[7] We used the R software environment for statistical computing (`http://www.r-project.org/`).

4.2 Evaluating the Concept Selection Indices

In the following we measure the performance of the three concept selection indices with respect to our data. The experimental setting is as follows:

We first select a number of N (1500) concepts with best selection index. The selected concepts are aligned with the classes translated from VerbNet (see Section 5): For each translated class, we select the concept with best precision/recall f-measure. Then we associate to the concept with best f-measure the thematic roles of the translated VN class. Next we compare the obtained ⟨verb, thematic role set⟩ associations with those given by a reference. As for our task recall is more important than precision, we use the $F2$ measure, which gives more weight to recall, for comparison.

As reference we use the data used for training the classifier for learning the translated VN classes (see Section 5): we are checking which index selects the most relevant concepts, that is those best matching the translated classes. The reference consists of the ⟨verb, semantic role set⟩ pairs marked as positive examples in the training set, ie. those for which we considered that the French verbs could have the semantic roles given by the English VN class. Table 2 shows

	cov.	**prec.**	**rec.**	**F2**
stab only	39.88	18.96	32.55	26.27
sep only	34.25	28.37	21.52	23.41
prob only	35.53	26.60	20.73	22.38
w/o filtering	100	12.30	60.96	26.30

Table 2: F2 scores and coverage for stability, separation and the 6th probability 10-quantile.

the F2 scores and coverage when using only one index at a time. For stability and separation we applied the method above on the top ranking 1500 concepts. Regarding probability, at first sight, we should consider best the concepts with lowest probability – because the probability of their intents of being closed by chance only is accordingly low. However, looking at the data we found that these concepts have very few verbs and large intent (frame) sets - which rather suggest improbable or rare verb groups. On the other hand, the interpretation of concept probability suggests that a concept with a probability close to 1 could occur by chance only. For these reasons, to assess probability separately we settled on the 6th 10 quantile. The results confirm the observations of Klimushkin et al. (2010): stability alone gives F2 scores close to an upper bound – the results obtained without filtering, ie. aligning the translated classes with all the concepts of the lattice. The results for separation and probability are several points lower.

As we only select $\sim 10\%$ of the total number of concepts we also have to make sure that the selected concepts cover at least a reasonable amount of verbs. The **cov** column gives the percentage of verbs in the lattice covered by the selected concepts. It shows that using only one index at a time the pre-selected concepts would contain only $35\% - 40\%$ of the verbs in the entire lattice, which is unsatisfactory.

Klimushkin et al. (2010) investigate the performance of the stability, separation and probability indices at finding the original concepts in lattices produced from contexts which were previously altered by introducing two types of noise: *Type I noise* is obtained by altering every cell in the context with some probability, *Type II noise* is obtained by adding a given number or proportion of random objects or attributes. According to this, our contexts are affected by Type I noise rather than Type II. Klimushkin et al. (2010) found that stability was most effective at sorting out Type II noise, but also proved helpful in the case of Type I noise. In contrast, they suggest that separation and probability can not be used on their own but should rather serve as a normalising measure for stability. The most promising combination seemed to be: $\text{stability} + k_{sep} \cdot \text{separation} - k_{prob} \cdot \text{probability}$.

In the following we start from the assumption that the most effective index for selecting relevant concepts is given by a linear combination of stability, separation and probability: $k_{stab} \cdot \text{stability} + k_{sep} \cdot \text{separation} - k_{prob} \cdot \text{probability}$, and empirically determine the coefficients k_{stab}, k_{sep} and k_{prob} such that the selected concepts perform best with respect to our task.

We proceed as follows: We choose k_{stab}, k_{sep} and k_{prob}. We then compute the corresponding linear combination for the concepts and select the 1500 concepts ranking highest. As in the previous experiments, we measure the relevance of the selected concepts by aligning the concepts with the translated VN classes and by comparing the alignments with the same reference as before. We consider the "best" $k_{stab}, k_{sep}, k_{prob}$ combination the one giving highest F2 scores and good coverage.

Table 3a shows the results for a first series of experiments where k_{stab} and k_{sep} were assigned the values 0.5 and 1 and k_{prob} 0.25 and 0.5 (The lines are sorted by decreasing F2 score). They suggest that the stability and separation coefficients had less impact on coverage and F2 score than the probability coefficient. Interestingly the coverage is correlated with the F2 score.

In the second series of experiments, shown in Table 3b, we kept the stability and separation coefficients fixed and varied only the probability coefficient. These results suggest that the probability coefficient may not help at selecting the most relevant concepts in our setting. This may be due first to the fact that our attributes are not independent (we assumed independence of attributes when setting up the formula for computing the probability index) and second to the fact that we had to approximate the probability index and this approximation may not be accurate enough.

In the next series of experiments we investigated the impact of the number of preselected concepts (500). The results showed that with this smaller number of concepts the selected concepts reached a slightly smaller F2 score but a substantially lower coverage. Also, in this configuration the probability index did seem to be helpful. Preselecting 1000 concepts confirmed the previously observed tendencies: The F2 score and coverage were only slightly lower than when preselecting 1500 concepts and again the probability index seemed to have only a small impact on the overall results.

(a) F2 and coverage when $k_{stab}, k_{sep} \in \{0.5, 1\}, k_{prob} \in \{0.25, 0.5\}$.

k_{stab}	k_{sep}	k_{prob}	**cov.**	**prec.**	**rec.**	**F2**
1	1	0.25	98.04	11.87	55.19	24.89
1	0.5	0.25	98.04	11.87	55.19	24.89
1	0.5	0.5	57.69	17.08	30.18	24.04
1	1	0.5	56.15	17.45	29.13	23.82
0.5	0.5	0.25	56.15	17.45	29.13	23.82
0.5	1	0.25	53.81	18.03	27.82	23.36
0.5	0.5	0.5	49.72	18.55	26.25	23.06
0.5	1	0.5	49.90	18.61	25.98	22.95

(b) F2 and coverage when k_{stab} and k_{sep} are kept fixed and k_{prob} varies.

k_{stab}	k_{sep}	k_{prob}	**cov.**	**prec.**	**rec.**	**F2**
1	1	0	98.04	12.05	55.12	25.16
1	1	0.05	98.04	12.05	55.12	25.16
1	1	0.005	98.04	12.05	55.12	25.16
1	1	0.0005	98.04	12.05	55.12	25.16
1	1	0.1	98.00	11.91	55.38	25.00
1	1	0.2	98.08	11.88	55.12	24.91
1	1	0.25	98.04	11.87	55.12	24.89
1	1	0.3	98.00	11.79	55.38	24.80
1	1	0.4	59.95	16.27	31.23	23.91
1	1	0.5	56.16	17.45	29.13	23.82
w/o filtering			100	12.30	60.96	26.30

Table 3: F2 scores and coverage for various $k_{stab}, k_{sep}, k_{prob}$ combinations.

From these experiments we conclude the following: First they suggest that the best linear combination is the sum of the stability and separation indices as the F2 measure and the coverage for this combination are similar to those of an upper bound, ie. the alignment obtained without filtering. They show that selecting only $\sim 10\%$ of the original lattice gives a verb, frame, semantic role set alignment which is close to the alignment obtained when using the entire lattice and that the pre-selected concepts also have a similar coverage.

Second, it does not seem evident that probability has a positive effect on the selected concepts. However, it does improve f-measure when the number of selected concepts is lower (500 or 1000 vs. 1500 in our experiments). Hence, for our application we concluded that it is a better strategy to select a larger number of concepts (1500) and not take probability into account. This is even more so as the probability index in our case should be taken with caution because first we had to use an approximation to compute it which may be too rough, and second the computation of probability is based on the independence of attributes which is not warranted in our case.

5 Associating French Verbs with Thematic Role Sets.

We associate French verbs with thematic role sets by translating the English VerbNet classes to French using 3 English-French dictionaries. In the following we first briefly describe the relevant resources, ie. VerbNet and the dictionaries before giving the translation methodology. As for this paper only the translated classes, but not the method to produce them is relevant[8] we only very briefly sketch the methodology.

[8] Of course better translated classes will result in a better performance of our method, but it is not straight forward to evaluate the quality of the translated classes.

VerbNet (Schuler (2006)) is the largest electronic verb classification for English. It was created manually and classifies 3626 verbs using 411 classes. Each VN class includes among other things a set of verbs, a set of subcategorisation frames and a set of thematic roles. Figure 2 shows an excerpt of the *amuse-31.1* class, with its verbs, thematic roles and subcategorisation frames.

verbs (242): abash, affect, afflict, amuse, annoy, ...
thematic roles: EXPERIENCER, CAUSE
frames (6):

NP V NP	EXPERIENCER V CAUSE
NP V ADV-Middle	EXPERIENCER V Adv
NP V NP-PRO-ARB	CAUSE V
...	

Fig. 2: Simplified VerbNet class *amuse-31.1*.

English-French dictionaries. We use the following resources to translate the verbs in the English VN classes to French: Sci-Fran-Euradic, a French-English bilingual dictionary, built and improved by linguists , Google dictionary[9] and Dicovalence van den Eynde and Mertens (2003)[10]. The merged dictionary contains 51242 French-English verb pairs.

In the following we describe our method for translating the English VerbNet classes to French.

The translation of VerbNet classes is bound to be very noisy because verbs are polysemous and the dictionaries typically give translations for several readings of the verb: Thus the dictionary may give several translations v_{fr} which do not correspond to the meaning given by the $\langle v_{en}, class\rangle$ pair or this meaning may even not be covered at all by the dictionary. To get more accurate translated VN classes we use a machine learning method, namely Support Vector Machines (SVM)[11]. We follow a straight forward SVM application scenario: we build all the French verb, VN class pairs $\langle v_{fr}, C_{VN}\rangle$ where v_{fr} is a translation of an English verb in C_{VN}. The classifier has to give a probability estimate about whether this association is correct or not.

For training the classifier we use the 160 verbs appearing in the gold standard proposed by Sun et al. (2010)[12]. We build the pairs $\langle v_{fr}, C_{VN}\rangle$ where v_{fr} is a verb in the gold standard which is a translation of a verb in C_{VN}. For each of these pairs we assessed whether or not there was a meaning of v_{fr} where the semantic roles involved in the event described by the verb were those given by C_{VN}. The features associated to the ⟨verb, class⟩ pairs are numeric and are extracted from the dictionaries and VerbNet.

[9] `http://www.google.com/dictionary`. We obtained 13824 French-English verb pairs.
[10] The number of French-English verb pairs we obtained is 11351
[11] We used *libsvm*, the software package and methodology presented on `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`, Chang and Lin (2011).
[12] In fact this is the only existing gold standard for French VerbNet style classes and we also use it for the overall evaluation of our system (not presented in this paper).

The trained classifier is then used to produce probability estimates for all verb, class instances. We select the 6000 pairs with highest probability estimates[13] and finally obtain the translated classes by assigning each verb in a selected pair to the corresponding class.

To give an idea of the quality of the obtained classes: The accuracy of the classifier on the held out test set was 90%, compared to a maximum accuracy of 93.84% for five fold cross-validation on the development set. The frequency distribution of the translated classes obtained this way is much closer to the distribution of verbs in VerbNet classes as when using an approach based only on translation frequencies, thus providing more accurate verb groups to guide the FCA concept - thematic roles associations.

6 The French Verb ↔ Thematic Role Sets ↔ Syntactic Frame Associations

As a detailed and thorough evaluation of the verb, thematic role sets and syntactic frames associations would be out of the scope of this paper we only give here an intuition of the type of information provided by our method. Following the preliminary investigations in the previous sections we associated French verbs with subcategorisation frames and thematic role sets according to the scheme listed below:

- We group the VerbNet thematic roles and assign to one class all the VN verbs whose class have the same role set. We then translate the obtained classes using the methods described in Section 5.
- We use FCA to group French verbs and syntactic frames associated to these verbs by the lexicons described in Section 3. The concept lattices we create are based on the formal contexts consisting of French verbs as objects and SCFs as attributes.
- We then select the 1500 concepts where the sum of the stability and separation indices is highest because in Section 4 we found this combination of concept selection indices to work best for our application.
- For each translated VN class we identify among the 1500 filtered FCA concepts the one(s) with best f-measure between precision and recall.

The translated VerbNet class is then associated with this FCA concept(s). Thus the verbs in the FCA concept are effectively associated with the thematic role set of the translated class and at the same time with the syntactic frames in the intent (attribute set) of the FCA concept. Figure 3 shows the associations between concepts, thematic role sets and frames generated by our method for some VN classes[14]. The figure shows the concepts associated to these thematic role sets and for each of these concepts: their attribute set (syntactic frames),

[13] In VerbNet there are 5726 verb, class pairs

[14] These are the classes occuring in the gold standard proposed by Sun et al. (2010), mentioned in Section 5.

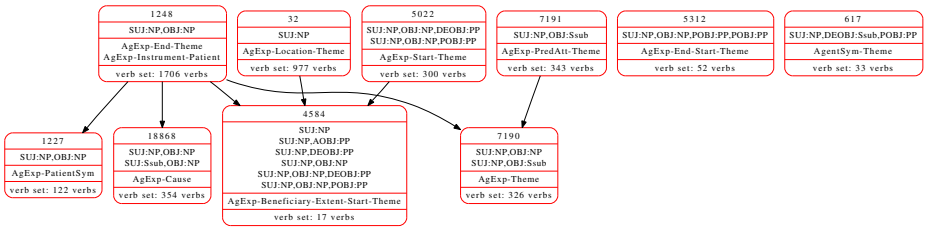

Fig. 3: French verb ↔ synt. frames ↔ thematic role set associations.

the associated thematic role set(s), the number of verbs in the concept and the hierarchical relations between the concepts as given by the concept lattice.

Thus for example the following 11 verbs (occuring in the gold standard) *bouger, déplacer, emporter, passer, promener, envoyer, expédier, jeter, porter, transmettre, transporter* are in concept 5312 and thereby may be used in the construction SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP[15] (according to our lexical resources). When they occur in this construction they are associated with the thematic role set *AgExp, End, Start, Theme*, i.e. the semantic roles involved are an AGENT or EXPERIENCER, a START point, an END point and a THEME. The listed verbs are all verbs of movement where an agent may move a theme from a start point to an end point – therefore in this case the associations with the syntactic frame and thematic role set seem to be correct. An NLP system which encounters the verb *déplacer* for example, used in the construction SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP could infer that possible thematic roles involved in the described event are an AGENT (or EXPERIENCER), a THEME, an END point and a START point. However, it still would not know which thematic role is realised by which syntactic argument.

There are also some problems with these associations. As can be seen in Figure 3, there is one case where the classification maps the same concept to two distinct VerbNet classes (*AgExp-End-Theme* and *AgExp-Instrument-Patient*). In addition, verbs in sub-concepts inherit the class label of the super-concept. Although there are verbs which belong to several VN classes, in many cases this multiple mapping was not warranted. Improving the precision of these mappings requires further investigations.

7 Conclusion

We introduced a new approach to verb clustering which involves the combined use of the English VerbNet, a bilingual English-French lexicon and a merged subcategorisation lexicon for French. Using these resources, we built two classifications, one derived from the English VN by translation and the other, from the subcategorisation lexicons via the construction of a formal concept lattice. We then use the translated VN to associate FCA concepts with VN classes

[15] a transitive construction with two additional prepositional objects

and thereby associate verbs with both syntactic frames and a thematic role set. We explored the performance of the concept selection indices introduced by Klimushkin et al. (2010) which are *stability*, *separation* and *probability* at selecting most relevant concepts with respect to our data and found that the sum of stability and separation gave best results in the setting of our application. These results were similar to those obtained without filtering, showing that this combination of the indices did indeed allow to select the most relevant concepts with respect to our data. Finally we showed the French verb, syntactic constructions and semantic role sets associations we obtained and briefly illustrated their potential use. Thus Formal Concept Analysis in combination with the concept selection indices, translation and set mapping methods proved an adequate method in this knowledge acquisition process.

Bibliography

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

Briscoe, T. and Carroll, J. (1993). Generalized probabilistic lr parsing of natural language (corpora) with unification-based grammars. *Comput. Linguist.*, 19(1):25–59.

Carroll, J. and Fang, A. C. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an hpsg parser. In *IJCNLP*, pages 646–654.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Cimiano, P., S.Staab, and Tane, J. (2003). Automatic Acquisition of Taxonomies from Text: FCA meets NLP. In *Proceedings of the PKDD/ECML'03 International Workshop on Adaptive Text Extraction and Mining (ATEM)*, pages 10–17.

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

Falk, I. and Gardent, C. (2010). Bootstrapping a Classification of French Verbs Using Formal Concept Analysis. In *Interdisciplinary Workshop on Verbs Interdisciplinary Workshop on Verbs*, page 6, Pisa Italy.

Falk, I., Gardent, C., and Lorenzo, A. (2010). Using Formal Concept Analysis to Acquire Knowledge about Verbs. In *Concept Lattices and their applications*, page 12, Sevilla, Spain.

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

Guillet, A. and Leclère, C. (1992). *La structure des phrases simples en français. 2 : Constructions transitives locatives*. Droz, Geneva.

Hadouche, F. and Lapalme, G. (2010). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages*, pages 193–220. Mise en page différente que celle parue dans la revue.

Jay, N., Kohler, F., and Napoli, A. (2008). Analysis of social communities with iceberg and stability-based concept lattices. In *ICFCA'08: Proceedings of the 6th international conference on Formal concept analysis*, pages 258–272, Berlin, Heidelberg. Springer-Verlag.

Klimushkin, M., Obiedkov, S., and Roth, C. (2010). Approaches to the selection of relevant concepts in the case of noisy data. In Kwuida, L. and Sertkaya, B., editors, *Formal Concept Analysis*, volume 5986 of *Lecture Notes in Computer Science*, chapter 18, pages 255–266. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Kupść, A. and Abeillé, A. (2008). Growing treelex. In Gelbkuh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin / Heidelberg.

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

Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

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

Priss, U. (2005). Linguistic Applications of Formal Concept Analysis. In Ganter, B., Stumme, G., and Wille, R., editors, *Formal Concept Analysis*, volume 3626 of *Lecture Notes in Computer Science*, pages 149–160–160. Springer Berlin / Heidelberg.

Roth, C., Obiedkov, S. A., and Kourie, D. G. (2006). Towards concise representation for taxonomies of epistemic communities. In *CLA*, pages 240–255.

Saint-Dizier, P. (1999). Alternation and verb semantic classes for french: Analysis and class formation. In *Predicative forms in natural language and in lexical knowledge bases*. Kluwer Academic Publishers.

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania.

Sporleder, C. (2002). A Galois Lattice based Approach to Lexical Inheritance Hierarchy Learning. In *15th European Conference on Artificial Intelligence (ECAI'02): Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, Lyon, France*.

Sun, L., Korhonen, A., Poibeau, T., and Messiant, C. (2010). Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1056–1064, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tolone, E. (2011). *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. PhD thesis, LIGM, Université Paris-Est, France, Laboratoire d'Informatique Gaspard-Monge, Université Paris-Est Marne-la-Vallée, France. (326 pp.).

van den Eynde, K. and Mertens, P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104.