Investigating the cross-linguistic potential of VerbNet -style classification

Lin Sun and Anna Korhonen
Computer Laboratory
University of Cambridge
ls418,alk23@cl.cam.ac.uk

Thierry Poibeau
LaTTiCe, UMR8094
CNRS & ENS
thierry.poibeau@ens.fr

Cédric Messiant
LIPN, UMR7030
CNRS & U. Paris 13
cedric.messiant@lipn.fr

Abstract

Verb classes which integrate a wide range of linguistic properties (Levin, 1993) have proved useful for natural language processing (NLP) applications. However, the real-world use of these classes has been limited because for most languages, no resources similar to VerbNet (Kipper-Schuler, 2005) are available. We apply a verb clustering approach developed for English to French – a language for which no such experiment has been conducted yet. Our investigation shows that not only the general methodology but also the best performing features are transferable between the languages, making it possible to learn useful VerbNet style classes for French automatically without language-specific tuning.

1 Introduction

A number of verb classifications have been built to support natural language processing (NLP) tasks (Grishman et al., 1994; Miller, 1995; Baker et al., 1998; Palmer et al., 2005; Kipper-Schuler, 2005; Hovy et al., 2006). These include both syntactic and semantic classifications, as well as ones which integrate aspects of both. Classifications which integrate a wide range of linguistic properties can be particularly useful for NLP applications suffering from data sparseness. One such classification is VerbNet (Kipper-Schuler, 2005). Building on the taxonomy of Levin (1993), VerbNet groups verbs (e.g. *deliver, post, dispatch*) into classes (e.g. SEND) on the basis of their shared meaning components and syntactic behaviour, identified in terms of meaning preserving diathesis alternations. Such classes can be identified across the entire lexicon, and they may also apply across languages, since their meaning components are said to be cross-linguistically applicable (Jackendoff, 1990).

Offering a powerful tool for generalization, abstraction and prediction, VerbNet classes have been used to support many important NLP tasks, including e.g. computational lexicography, parsing, word sense disambiguation, semantic role labeling, information extraction, question-answering, and machine translation (Swier and Stevenson, 2004; Dang, 2004; Shi and Mihalcea, 2005; Abend et al., 2008). However, to date their exploitation has been limited because for most languages, no Levin style classification is available.

Since manual classification is costly (Kipper et al., 2008) automatic approaches have been proposed recently which could be used to learn novel classifications in a cost-effective manner (Joanis et al., 2008; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008; Vlachos et al., 2009; Sun and Korhonen, 2009). However, most work on Levin type classification has focussed on English. Large-scale research on other languages such as German (Schulte im Walde, 2006) and Japanese (Suzuki and Fukumoto, 2009) has focussed on semantic classification. Although the two classification systems have shared properties, studies comparing the overlap between VerbNet and WordNet (Miller, 1995) have reported that the mapping is only partial and many to many due to fine-grained nature of classes based on synonymy (Shi and Mihalcea, 2005; Abend et al., 2008).

Only few studies have been conducted on Levin style classification for languages other than English. In their experiment involving 59 verbs and three classes, Merlo et al. (2002) applied a supervised approach developed for English to Italian, obtaining high accuracy (86.3%). In another experiment with 60 verbs and three classes,

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1056–1064,
Beijing, August 2010

they showed that features extracted from Chinese translations of English verbs can improve English classification. These results are promising, but those from a later experiment by Ferrer (2004) are not. Ferrer applied a clustering approach developed for English to Spanish, and evaluated it against the manual classification of Vázquez et al. (2000), constructed using criteria similar (but not identical) to Levin's. This experiment involving 514 verbs and 31 classes produced results only slightly better than the random baseline.

In this paper, we investigate the cross-linguistic potential of Levin style classification further. In past years, verb classification techniques – in particular unsupervised ones – have improved considerably, making investigations for a new language more feasible. We take a recent verb clustering approach developed for English (Sun and Korhonen, 2009) and apply it to French – a major language for which no such experiment has been conducted yet. Basic NLP resources (corpora, taggers, parsers and subcategorization acquisition systems) are now sufficiently developed for this language for the application of a state-of-the-art verb clustering approach to be realistic.

Our investigation reveals similarities between the English and French classifications, supporting the linguistic hypothesis (Jackendoff, 1990) and the earlier result of Merlo et al. (2002) that Levin classes have a strong cross-linguistic basis. Not only the general methodology but also best performing features are transferable between the languages, making it possible to learn useful classes for French automatically without language-specific tuning.

2 French Gold Standard

The development of an automatic verb classification approach requires at least an initial gold standard. Some syntactic (Gross, 1975) and semantic (Vossen, 1998) verb classifications exist for French, along with ones which integrate aspects of both (Saint-Dizier, 1998). Since none of these resources offer classes similar to Levins', we followed the idea of Merlo et al. (2002) and translated a number of Levin classes from English to French. As our aim was to to investigate the cross-linguistic applicability of classes, we took an English gold standard which has been used to evaluate several recent clustering works – that of Sun et al. (2008). This resource includes 17 fine-grained Levin classes. Each class has 12 member verbs whose predominant sense in English (according to WordNet) belongs to that class.

Member verbs were first translated to French. Where several relevant translations were identified, each of them was considered. For each candidate verb, subcategorization frames (SCFs) were identified and diathesis alternations were considered using the criteria of Levin (1993): alternations must result in the same or extended verb sense. Only verbs sharing diathesis alternations were kept in the class.

For example, the gold standard class 31.1 AMUSE includes the following English verbs: *stimulate, threaten, shock, confuse, upset, overwhelm, scare, disappoint, delight, exhaust, intimidate* and *frighten*. Relevant French translations were identified for all of them: *abattre, accabler, briser, déprimer, consterner, anéantir, épuiser, exténuer, écraser, ennuyer, éreinter, inonder*. The majority of these verbs take similar SCFs and diathesis alternations, e.g. *Cette affaire écrase Marie (de chagrin), Marie est écrasée par le chagrin, Le chagrin écrase Marie*. However, *stimuler* (*stimulate*) and *menacer* (*threaten*) do not, and they were therefore removed.

40% of translations were discarded from classes because they did not share the same alernations. The final version of the gold standard (shown in table 1) includes 171 verbs in 16 classes. Each class is named according to the original Levin class. The smallest class (30.3) includes 7 verbs and the largest (37.3) 16. The average number of verbs per class is 10.7.

3 Verb Clustering

We performed an experiment where we

- took a French corpus and a SCF lexicon automatically extracted from that corpus,
- extracted from these resources a range of features (lexical, syntactic and semantic) – a representative sample of those employed in recent English experiments,

Class No	Class	Verbs
9.1	PUT	accrocher, déposer, mettre, placer, répartir, réintégrer, empiler, emporter, enfermer, insérer, installer
10.1	REMOVE	ôter, enlever, retirer, supprimer, retrancher, débarrasser, soustraire, décompter, éliminer
11.1	SEND	envoyer, lancer, transmettre, adresser, porter, expédier, transporter, jeter, renvoyer, livrer
13.5.1	GET	acheter, prendre, saisir, réserver, conserver, garder, préserver, maintenir, retenir, louer, affréter
18.1	HIT	cogner, heurter, battre, frapper, fouetter, taper, rosser, brutaliser, éreinter, maltraiter, corriger,
22.2	AMALGAMATE	incorporer, associer, réunir, mélanger, mêler, unir, assembler, combiner, lier, fusionner
29.2	CHARACTERIZE	appréhender, concevoir, considérer, décrire, définir, dépeindre, désigner, envisager, identifier, montrer, percevoir, représenter, ressentir
30.3	PEER	regarder, écouter, examiner, considérer, voir, scruter, dévisager
31.1	AMUSE	abattre, accabler, briser, déprimer, consterner, anéantir, épuiser, exténuer, écraser, ennuyer, éreinter, inonder,
36.1	CORRESPOND	coopérer, participer, collaborer, concourir, contribuer, prendre part, s'associer, travaille
37.3	MANNER OF SPEAKING	râler, gronder, crier, ronchonner, grogner, bougonner, maugréer, rouspéter, grommeler, larmoyer, gémir, geindre, hurler, gueuler, brailler, chuchoter
37.7	SAY	dire, révéler, déclarer, signaler, indiquer, montrer, annoncer, répondre, affirmer, certifier, répliquer
43.1	LIGHT EMISSION	briller, étinceler, flamboyer, luire, resplendir, pétiller, rutiler, rayonner., scintiller
45.4	CHANGE OF STATE	mélanger, fusionner, consolider, renforcer, fortifier, adoucir, polir, atténuer, tempérer, pétrir, façonner, former
47.3	MODES OF BEING	trembler, frémir, osciller, vaciller, vibrer, tressaillir, frissonner, palpiter, grésiller, trembloter, palpiter
51.3.2	RUN	voyager, aller, se promener, errer, circuler, se déplacer, courir, bouger, naviguer, passer

Table 1: A Levin style gold standard for French

- clustered the features using a method which has proved promising in both English and German experiments: spectral clustering,
- evaluated the clusters both quantitatively (using the gold standard) and qualitatively,
- and compared the performance to that recently obtained for English in order to gain a better understanding of the cross-linguistic and language-specific properties of verb classification

This work is described in the subsequent sections.

3.1 Data: the LexSchem Lexicon

We extracted the features for clustering from LexSchem (Messiant et al., 2008). This large subcategorization lexicon provides SCF frequency information for 3,297 French verbs. It was acquired fully automatically from Le Monde newspaper corpus (200M words from years 1991-2000) using ASSCI – a recent subcategorization acquisition system for French (Messiant, 2008). Systems similar to ASSCI have been used in recent verb classification works e.g. (Schulte im Walde, 2006; Li and Brew, 2008; Sun and Korhonen, 2009). Like these other systems, ASSCI takes raw corpus data as input. The data is first tagged and lemmatized using the Tree-Tagger and then parsed using Syntex (Bourigault et al., 2005). Syntex is a shallow parser which employs a combination of statistics and heuristics to identify grammatical relations (GRs) in sentences. ASSCI considers GRs where the target verbs occur and constructs SCFs from nominal, prepositional and adjectival phrases, and infinitival and subordinate clauses. When a verb has no dependency, its SCF is considered as intransitive. ASSCI assumes no predefined list of SCFs but almost any combination of permitted constructions can appear as a candidate SCF. The number of automatically generated SCF types in LexSchem is 336.

Many candidate SCFs are noisy due to processing errors and the difficulty of argument-adjunct distinction. Most SCF systems assume that true arguments occur in argument positions more frequently than adjuncts. Many systems also integrate filters for removing noise from system output. When LexSchem was evaluated after filter-

ing its F-measure was 69 – which is similar to that of other current SCF systems (Messiant et al., 2008) We used the unfiltered version of the lexicon because English experiments have shown that information about adjuncts can help verb clustering (Sun et al., 2008).

4 Features

Lexical entries in LexSchem provide a variety of material for verb clustering. Using this material, we constructed a range of features for experimentation. The first three include basic information about SCFs:

F1: SCFs and their relative frequencies with individual verbs. SCFs abstract over particles and prepositions.

F2: F1, with SCFs parameterized for the tense (the POS tag) of the verb.

F3: F2, with SCFs parameterized for prepositions (PP).

The following six features include information about the lexical context (co-occurrences) of verbs. We adopt the best method of Li and Brew (2008) where collocations (COs) are extracted from the window of words immediately preceding and following a lemmatized verb. Stop words are removed prior to extraction.

F4, F6, F8: COs are extracted from the window of 4, 6 and 8 words, respectively. The relative word position is ignored.

F5, F7, F9: F4, F6 and F8 with the relative word position recorded.

The next four features include information about lexical preferences (LP) of verbs in argument head positions of specific GRs associated with the verb:

F10: LP(PREP): the type and frequency of prepositions in the preposition (PREP) relation.

F11: LP(SUBJ): the type and frequency of nouns in the subject (SUBJ) relation.

F12: LP(IOBJ): the type and frequency of nouns in the object (OBJ) and indirect object (IOBJ) relation.

F13: LP(ALL): the combination of F10-F13.

The final two features refine SCF features with LPs and semantic information about verb selectional preferences (SP):

F14-F16: F1-F3 parameterized for LPs.

F17: F3 refined with SPs.

We adopt a fully unsupervised approach to SP acquisition using the method of Sun and Korhonen (2009), with the difference that we determine the optimal number of SP clusters automatically following Zelnik-Manor and Perona (2004). The method is introduced in the following section. The approach involves (i) taking the GRs (SUBJ, OBJ, IOBJ) associated with verbs, (ii) extracting all the argument heads in these GRs, and (iii) clustering the resulting N most frequent argument heads into M classes. The empirically determined N 200 was used. The method produced 40 SP clusters.

5 Clustering Methods

Spectral clustering (SPEC) has proved promising in previous verb clustering experiments (Brew and Schulte im Walde, 2002; Sun and Korhonen, 2009) and other similar NLP tasks involving high dimensional feature space (Chen et al., 2006). Following Sun and Korhonen (2009) we used the MNCut spectral clustering (Meila and Shi, 2001) which has a wide applicability and a clear probabilistic interpretation (von Luxburg, 2007; Verma and Meila, 2005). However, we extended the method to determine the optimal number of clusters automatically using the technique proposed by (Zelnik-Manor and Perona, 2004).

Clustering groups a given set of verbs $V = \{v_n\}_{n=1}^{N}$ into a disjoint partition of K classes. SPEC takes a similarity matrix as input. All our features can be viewed as probabilistic distributions because the combination of different features is performed via parameterization. Thus we use the Jensen-Shannon divergence (JSD) to construct the similarity matrix. The JSD between

two feature vectors v and v' is $d_{jsd}(v, v') = \frac{1}{2}D(v||m) + \frac{1}{2}D(v'||m)$ where D is the Kullback-Leibler divergence, and m is the average of the v and v'.

The similarity matrix W is constructed where $W_{ij} = \exp(-d_{jsd}(v, v'))$. In SPEC, the similarities W_{ij} are viewed as the connection weight ij of a graph G over V. The similarity matrix W is thus the adjacency matrix for G. The degree of a vertex i is $d_i = \sum_{j=1}^{N} w_{ij}$. A cut between two partitions A and A' is defined to be $\text{Cut}(A, A') = \sum_{m \in A, n \in A'} W_{mn}$.

The similarity matrix W is normalized into a stochastic matrix P.

$$P = D^{-1}W \quad (1)$$

The degree matrix D is a diagonal matrix where $D_{ii} = d_i$.

It was shown by Meila and Shi (2001) that if P has the K leading eigenvectors that are piecewise constant[1] with respect to a partition I^* and their eigenvalues are not zero, then I^* minimizes the multiway normalized cut(MNCut):

$$\text{MNCut}(I) = K - \sum_{k=1}^{K} \frac{\text{Cut}(I_k, I_k)}{\text{Cut}(I_k, I)}$$

P_{mn} can be interpreted as the transition probability between vertices m, n. The criterion can thus be expressed as $\text{MNCut}(I) = \sum_{k=1}^{K}(1 - P(I_k \rightarrow I_k|I_k))$ (Meila, 2001), which is the sum of transition probabilities across different clusters. This criterion finds the partition where the random walks are most likely to happen within the same cluster. In practice, the leading eigenvectors of P are not piecewise constant. But we can extract the partition by finding the approximately equal elements in the eigenvectors using a clustering algorithm like K-Means.

As the value of K is not known beforehand, we use Zelnik-Manor and Perona (2004)'s method to estimate it. This method finds the optimal value by minimizing a cost function based on the eigenvector structure of W.

Like Brew and Schulte im Walde (2002), we compare SPEC against a K-Means baseline. We used the Matlab implementation with euclidean distance as the distance measure.

[1]The eigenvector v is piecewise constant with respect to I if $v(i) = v(j) \forall i, j \in I_k$ and $k \in 1, 2...K$

6 Experimental Evaluation

6.1 Data and Pre-processing

The SCF-based features (F1-F3 and F14-F17) were extracted directly from LexSchem. The CO (F4-F9) and LP features (F10-F13) were extracted from the raw and parsed corpus sentences, respectively, which were used for creating the lexicon. Features that only appeared once were removed. Feature vectors were normalized by the sum of the feature values before clustering. Since our clustering algorithms have an element of randomness, we repeated clustering multiple times. We report the results that minimize the distortion (the distance to cluster centroid).

6.2 Evaluation Measures

We employ the same measures for evaluation as previously employed e.g. by Ó Séaghdha and Copestake (2008) and Sun and Korhonen (2009).

The first measure is modified purity (mPUR) – a global measure which evaluates the mean precision of clusters. Each cluster is associated with its prevalent class. The number of verbs in a cluster K that take this class is denoted by $n_{prevalent}$(K). Verbs that do not take it are considered as errors. Clusters where $n_{prevalent}$(K) = 1 are disregarded as not to introduce a bias towards singletons:

$$\text{mPUR} = \frac{\sum_{n_{prevalent(k_i)>2}} n_{prevalent(k_i)}}{\text{number of verbs}}$$

The second measure is weighted class accuracy (ACC): the proportion of members of dominant clusters DOM-CLUST$_i$ within all classes c_i.

$$\text{ACC} = \frac{\sum_{i=1}^{C} \text{verbs in DOM-CLUST}_i}{\text{number of verbs}}$$

mPUR and ACC can be seen as a measure of precision(P) and recall(R) respectively. We calculate F measure as the harmonic mean of P and R:

$$\text{F} = \frac{2 \cdot \text{mPUR} \cdot \text{ACC}}{\text{mPUR} + \text{ACC}}$$

The random baseline (BL) is calculated as follows: $\text{BL} = 1/\text{number of classes}$

7 Evaluation

7.1 Quantitative Evaluation

In our first experiment, we evaluated 116 verbs – those which appeared in LexSchem the minimum

of 150 times. We did this because English experiments had shown that due to the Zipfian nature of SCF distributions, 150 corpus occurrences are typically needed to obtain a sufficient number of frames for clustering (Sun et al., 2008).

Table 2 shows F-measure results for all the features. The 4th column of the table shows, for comparison, the results of Sun and Korhonen (2009) obtained for English when they used the same features as us, clustered them using SPEC, and evaluated them against the English version of our gold standard, also using F-measure[2].

As expected, SPEC (the 2nd column) outperforms K-Means (the 3rd column). Looking at the basic SCF features F1-F3, we can see that they perform significantly better than the BL method. F3 performs the best among the three features both in French (50.6 F) and in English (63.3 F). We therefore use F3 as the SCF feature in F14-F17 (the same was done for English).

In French, most CO features (F4-F9) outperform SCF features. The best result is obtained with F7: 55.1 F. This is clearly better than the best SCF result 50.6 (F3). This result is interesting since SCFs correspond better than COs with features used in manual Levin classification. Also, SCFs perform considerably better than COs in the English experiment (we only have the result for F4 available, but it is considerably lower than the result for F3). However, earlier English studies have reported contradictory results (e.g. Li and Brew (2008) showed that CO performs better than SCF in supervised verb classification), indicating that the role of CO features in verb classification requires further investigation.

Looking at the LP features, F13 produces the best F (52.7) for French which is slightly better than the best SCF result for the language. Also in English, F13 performs the best in this feature group and yields a higher result than the best SCF-based feature F3.

Parameterizing the best SCF feature F3 with LPs (F14-16) and SPs (F17) yields better performance in French. F15 and F17 have the F of 54.5 and 54.6, respectively. These results are so close to the result of the best CO feature F7 (55.1 – which is the highest result in this experiment) that the differences are not statistically significant. In English, the results of F14-F17 are similarly good; however, only F17 beats the already high performance of F13.

On the basis of this experiment, it is difficult to tell whether shallow CO features or more sophisticated SCF-based features are better for French. In the English experiment sophisticated features performed better (the SCF-SP feature was the best). However, the English experiment employed a much larger dataset. These more sophisticated features may suffer from data sparseness in our French experiment since although we required the minimum of 150 occurrences per verb, verb clustering performance tends to improve when more data is available, and given the fine-grained nature of LexShem SCFs it is likely that more data is required for optimal performance.

We therefore performed another experiment with French on the full set of 147 verbs, using SPEC, where we investigated the effect of instance filtering on the performance of the best features from each feature group: F3, F7, F13 and F17. The results shown in Table 3 reveal that the performance of the features remains fairly similar until the instance threshold of 1000. When 2000 occurrences per verb are used, the differences become clearer, until at the threshold of 4000, it is obvious that the most sophisticated SCF-SP feature F17 is by far the best feature for French (65.4 F) and the SCF feature F3 the second best (60.5 F). The CO-feature F7 and the LP feature F13 are not nearly as good (53.4 and 51.0 F).

Although the results at different thresholds are not comparable due to the different number of verbs and classes (see columns 2-3), the results for features at the same threshold are. Those results suggest that when 2000 or more occurrences per verb are used, most features perform like they performed for English in the experiment of Sun and Korhonen (2009), with CO being the least informative[3] and SCF-SP being the most informa-

[2]Note that the results for the two languages are not mutually comparable due to differences in test sets, data sizes, and feature extraction systems (see Section 8 for discussion). The results for English are included so that we can compare the relative performance of individual features in the two languages in question.

[3]However, it is worth noting that CO is not a useless feature. As table 3 shows, when 150 or fewer occurrences are

		SPEC	K	Eng.
BL		6.7	6.7	6.7
F1	SCF	42.4	39.3	57.8
F2	SCF(POS)	45.9	40.3	46.7
F3	SCF(PP)	**50.6**	36.9	63.3
F4	CO(4)	50.3	38.2	40.9
F5	CO(4+loc)	48.8	26.3	-
F6	CO(6)	52.7	29.2	-
F7	CO(6+loc)	**55.1**	33.8	-
F8	CO(8)	54.2	36.4	-
F9	CO(8+loc)	54.6	37.2	-
F10	LP(PREP)	35.5	32.8	49.0
F11	LP(SUBJ)	33.7	23.6	-
F12	LP(OBJ)	50.1	33.3	-
F13	LP(ALL)	**52.7**	40.1	74.6
F14	SCF+LP(SUBJ)	50.3	40.1	71.7
F15	SCF+LP(OBJ)	**54.5**	35.6	74.0
F16	SCF+LP(SUBJ+OBJ)	53.4	36.2	73.0
F17	SCF+SP	54.6	39.8	80.4

Table 2: Results for all the features for French (SPEC and K-means) and English (SPEC)

THR	Verbs	Cls	F3	F7	F13	F17
0	147	15	43.7	57.5	43.3	50.1
50	137	15	47.9	56.1	44.8	49.1
100	125	15	49.2	54.3	44.8	49.5
150	116	15	50.6	55.1	52.7	54.6
200	110	15	54.9	52.9	49.7	52.5
400	96	15	52.7	52.9	43.9	53.2
1000	71	15	51.4	54.0	44.8	54.5
2000	59	12	52.3	45.9	42.7	53.5
3000	51	12	55.7	49.0	46.8	59.2
4000	43	10	60.5	53.4	51.0	**65.4**

Table 3: The effect of verb frequency

tive feature. The only exception is the LP feature which performed better than CO in English.

7.2 Qualitative Evaluation

We conducted qualitative analysis of the clusters for French: those created using SPEC with F17 and F3. Verbs in the gold standard classes 29.2, 36.1, 37.3, 37.7 and 47.3 (Table 1) performed particularly well, with the majority of member verbs found in the same cluster. These verbs are ideal for clustering because they have distinctive syntactic-semantic characteristics. For example, verbs in 29.2 CHARACTERIZE class (e.g. *concevoir, considérer, dépeindre*) not only have a very specific meaning but they also take high frequency SCFs involving the preposition *comme* (Eng. *as*) which is not typical to many other classes. Interestingly, Levin classes 29.2, 36.1, 37.3, and 37.7 were among the best performing classes also in the supervised verb classification experiment of Sun et al. (2008) because these classes have distinctive characteristics also in English.

The benefit of sophisticated features which integrate also semantic (SP) information (F17) is particularly evident for classes with non-distinctive syntactic characteristics. For example, the intransitive verbs in 43.1 LIGHT EMISSION class (e.g. *briller, étinceler, flamboyer*) are difficult to cluster based on syntax only, but semantic features work because the verbs pose strong SPs on their subjects (entities capable of light emission). In the experiment of Sun et al. (2008), 43.1 was the worst performing class, possibly because no semantic features were used in the experiment.

The most frequent source of error is syntactic idiosyncracy. This is particularly evident for classes 10.1 REMOVE and 45.4 CHANGE OF STATE. Although verbs in these classes can take similar SCFs and alternations, only some of them are frequent in data. For example, the SCF *ôter X à Y* is frequent for verbs in 10.1, but not *ôter X de Y*. Although class 10.1 did not suffer from this problem in the English experiment of Sun et al. (2008), class 45.4 did. Class 45.4 performs particularly bad in French also because its member verbs are low in frequency.

Some errors are due to polysemy, caused partly by the fact that the French version of the gold standard was not controlled for this factor. Some verbs have their predominant senses in classes which are missing in the gold standard, e.g. the most frequent sense of *retenir* is *memorize*, not *keep* as in the gold standard class 13.5.1. GET.

Finally, some errors are not true errors but demonstrate the capability of clustering to learn novel information. For example, the CHANGE OF STATE class 45.4 includes many antonyms (e.g. *weaken* vs. *strenghten*). Clustering (using F17) separates these antonyms, so that verbs *adoucir, atténuer* and *tempérer* appear in one cluster and *consolider* and *renforcer* in another. Although these verbs share the same alternations, their SPs are different. The opposite effect can be observed when clustering maps together classes

available for a verb, CO outperforms all the other features in French, compensating for data sparseness.

which are semantically and syntactically related (e.g. 36.1 CORRESPOND and 37.7 SPEAK). Such classes are distinct in Levin and VerbNet, although should ideally be related. Cases such as these show the potential of clustering in discovering novel valuable information in data.

8 Discussion and Conclusion

When sufficient corpus data is available, there is a strong correlation between the types of features which perform the best in English and French. When the best features are used, many individual Levin classes have similar performance in the two languages. Due to differences in data sets direct comparison of performance figures for English and French is not possible. When considering the general level of performance, our best performance for French (65.4 F) is lower than the best performance for English in the experiment of Sun and Korhonen (2009). However, it does compare favourably to the performance of other state-of-the-art (even supervised) English systems (Joanis et al., 2008; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008; Vlachos et al., 2009). This is impressive considering that we experimented with a fully unsupervised approach originally developed for another language.

When aiming to improve performance further, employing larger data is critical. Most recent experiments on English have employed bigger data sets, and unlike us, some of them have only considered the predominant senses of medium-high frequency verbs. As seen in section 7.1, such differences in data can have significant impact on performance. However, parser and feature extraction performance can also play a big role in overall accuracy, and should therefore be investigated further (Sun and Korhonen, 2009). The relatively low performance of basic LP features in French suggests that at least some of the current errors are due to parsing. Future research should investigate the source of error at different stages of processing. In addition, it would be interesting to investigate whether language-specific tuning (e.g. using language specific features such as auxiliary classes) can further improve performance on French.

Earlier works most closely related to ours are those of Merlo et al. (2002) and Ferrer (2004). Our results contrast with those of Ferrer who showed that a clustering approach does not transfer well from English to Spanish. However, she used basic SCF and named entity features only, and a clustering algorithm less suitable for high dimensional data. Like us, Merlo et al. (2002) created a gold standard by translating Levin classes to another language (Italian). They also applied a method developed for English to Italian, and reported good overall performance using features developed for English. Although the experiment was small (focussing on three classes and a few features only) and involved supervised classification, the results agree with ours.

These experiments support the linguistic hypothesis that Levin style classification can be cross-linguistically applicable. A clustering technique such as the one presented here could be used as a tool for investigating whether classifications are similar across a wider range of more diverse languages. From the NLP perspective, the fact that an unsupervised technique developed for one language can be applied to another language without the need for substantial tuning means that automatic techniques could be used to hypothesise useful Levin style classes for further languages. This, in turn, could facilitate the creation of multilingual VerbNets in the future.

9 Acknowledgement

Our work was funded by the Royal Society University Research Fellowship (AK), the Dorothy Hodgkin Postgraduate Award (LS), the EPSRC grants EP/F030061/1 and EP/G051070/1 (UK) and the EU FP7 project 'PANACEA'.

References

Omri Abend, Roi Reichart, and Ari Rappoport. A supervised algorithm for verb disambiguation into VerbNet classes. In *Proc. of COLING*, pages 9–16, 2008.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *COLING-ACL*, pages 86–90, 1998.

Didier Bourigault, Marie-Paule Jacques, Cécile Fabre, Cécile Frérot, and Sylwia Ozdowska. Syntex, analyseur syntaxique de corpus. In *Actes des*

12èmes journées sur le Traitement Automatique des Langues Naturelles, 2005.

Chris Brew and Sabine Schulte im Walde. Spectral clustering for German verbs. In *Proc. of EMNLP*, pages 117–124, 2002.

Jinxiu Chen, Dong-Hong Ji, Chew Lim Tan, and Zheng-Yu Niu. Unsupervised relation disambiguation using spectral clustering. In *Proc. of COLING/ACL*, pages 89–96, 2006.

Hoa Trang Dang. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. PhD thesis, CIS, University of Pennsylvania, 2004.

Eva Esteve Ferrer. Towards a semantic classification of Spanish verbs based on subcategorisation information. In *Proc. of ACL Student Research Workshop*, 2004.

Ralph Grishman, Catherine Macleod, and Adam Meyers. Comlex syntax: building a computational lexicon. In *Proc. of COLING*, pages 268–272, 1994.

Maurice Gross. *Méthodes en syntaxe*. Hermann, Paris, 1975.

Eduard Hovy, Mitch Marcus, Martha Palmer, L. Ramshaw, and R. Weischedel. Ontonotes: The 90% solution. In *HLT/NAACL*, 2006.

Ray Jackendoff. *Semantic Structures*. The MIT Press, Cambridge, MA, 1990.

Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Nat. Lang. Eng.*, 14(3):337–367, 2008.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42:21–40, 2008.

Karin Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, PA, 2005.

Beth. Levin. English verb classes and alternations: A preliminary investigation. *Chicago, IL*, 1993.

Jianguo Li and Chris Brew. Which Are the Best Features for Automatic Verb Classification. In *Proc. of ACL*, pages 434–442, 2008.

Marina. Meila. The multicut lemma. Technical report, University of Washington, 2001.

Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. In *AISTATS*, 2001.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. A multilingual paradigm for automatic verb classification. In *Proc. of ACL*, 2002.

Cédric Messiant. ASSCI : A subcategorization frames acquisition system for French. In *Proc. of ACL Student Research Workshop*, pages 55–60, 2008.

Cédric Messiant, Thierry Poibeau, and Anna Korhonen. LexSchem: a Large Subcategorization Lexicon for French Verbs. In *Proc. of LREC*, 2008.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 1995.

Diarmuid Ó Séaghdha and Ann Copestake. Semantic classification with distributional kernels. In *Proc. of COLING*, pages 649–656, 2008.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 3(1):71–106, 2005.

Patrick Saint-Dizier. Verb Semantic Classes Based on 'alternations' and WordNet-like criteria . In P. Saint-Dizier, editor, *Predicative Forms in Natural language and lexical Knowledge Bases* , pages 247–279. Kluwer Academic, 1998.

Sabine Schulte im Walde. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 2006.

Lei Shi and Rada Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proc. of CICLing*, pages 100–111, 2005.

Lin Sun and Anna Korhonen. Improving verb clustering with automatically acquired selectional preferences. In *Proc. of EMNLP*, pages 638–647, 2009.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. Verb class discovery from rich syntactic data. *LNCS*, 4919:16, 2008.

Yoshimi Suzuki and Fumiyo Fukumoto. Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering. In *Proc. of TextGraphs-4*, pages 32–40, 2009.

Robert Swier and Suzanne Stevenson. Unsupervised semantic role labelling. In *Proc. of EMNLP*, 2004.

Gloria Vázquez, Ana Fernández, Irene Castellón, and M. Antonia Martí. Clasificación verbal: Alternancias de diátesis. In *Quaderns de Sintagma*. Universitat de Lleida, 2000.

Deepak Verma and Marina Meila. A comparison of spectral clustering algorithms. Technical report, Department of CSE University of Washington Seattle, 2005.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering. In *Proc. of the Workshop on on GEMS*, pages 74–82, 2009.

Ulrike von Luxburg. A tutorial on spectral clustering. *STAT COMPUT*, 17:395 – 416, 2007.

Piek Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *NIPS*, 17(1601-1608):16, 2004.