Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora

Alessandro Lenci[1], Barbara McGillivray[1], Simonetta Montemagni[2], Vito Pirrelli[2]
[1] Dipartimento di Linguistica, Università di Pisa, Pisa (Italy)
2 Istituto di Linguistica Computazionale, CNR, Pisa (Italy)
E-mail: alessandro.lenci@ilc.cnr.it, Barbara.Mcgillivray@aksis.uib.no, simonetta.montemagni@ilc.cnr.it, vito.pirrelli@ilc.cnr.it

Abstract

In this paper, we reported experiments of unsupervised automatic acquisition of Italian and English verb subcategorization frames (SCFs) from general and domain corpora. The proposed technique operates on syntactically shallow-parsed corpora on the basis of a limited number of search heuristics not relying on any previous lexico-syntactic knowledge about SCFs. Although preliminary, reported results are in line with state-of-the-art lexical acquisition systems. The issue of whether verbs sharing similar SCFs distributions happen to share similar semantic properties as well was also explored by clustering verbs that share frames with the same distribution using the Minimum Description Length Principle (MDL). First experiments in this direction were carried out on Italian verbs with encouraging results.

1. Introduction and State of the Art

Over the last decades, fully automatic acquisition of subcategorization frames (henceforth, SCFs) for English from large corpora has been pursued through a variety of different approaches. Brent & Berwick's (1991) system detects five SCFs by looking for attested contexts where argument slots are filled by closed-class lexical items (pronouns or proper names). Ushioda et al. (1993) use a finite-state NP parser to identify on a PoS tagged corpus six types of SCFs. Briscoe & Carroll (1997) extend to 163 the number of identified SCFs. Their system is able to build a SCF lexicon, whose entries include the relative frequency of the SCF classes. Potential SCF patterns are extracted from a dependency-based parsed corpus, and then filtered by hypothesis testing on binomial frequency data. Korhonen (2002) refines Briscoe and Carroll's system using back-off estimates on the WordNet semantic class of the verb's predominant sense, assuming that semantically similar verbs are also similar from a SCF point of view (see Levin's (1993) taxonomy of English verbs).

Most of these approaches presuppose a battery of predefined frames. SCFs acquisition is modelled as the task of detecting a verb's most likely SCFs within all its annotated syntactic contexts. This method has the serious shortcoming of requiring a priori specification of the SCFs to be detected. The negative effects of this assumption are particularly critical in languages for which no such SCF repertoires are already available. In such cases, a viable alternative is to model the acquisition process as a "SCF discovery" process in corpora. Basili et al. (1997) present a method for corpus-driven acquisition of subcategorization structures in Italian domain corpora. Starting from a parsed corpus, a conceptual clustering method is used to detect the different senses of each verb. Syntactic frames are then associated with the verbal senses given by the clusters to build a subcategorization lexicon. Zeman & Sarkar (2000) start from the Czech dependency treebank and apply machine learning techniques to learn associations between verbs and possible SCFs, taking as input a training corpus containing a list of verbs and their observed frames. Alonso et al. (2007) describe a method that assigns SCFs to unseen Spanish verbs making use of a syntactically and semantically annotated corpus. Bourigault & Frérot (2005) apply linguistic rules to identify the potential governors of a preposition; when dealing with ambiguous cases, they assign SCF probabilities obtained from previously seen non-ambiguous cases.

The work presented here [1] applies a variation of the "discovery approach" to SCF acquisition of Italian verbs. In a nutshell, our method simply requires a syntactically shallow-parsed corpus and a limited number of search heuristics, that do not rely on any previous knowledge about SCFs. This strategy has the main advantage of not presupposing any strict definition of SCF structure nor the a priori distinction between subcategorized arguments and optional adjuncts. In fact, our approach adheres to a looser notion of SCF including typical verb modifiers along with strongly selected arguments. This feature is particularly important when dealing with texts belonging to specialised domains: this is the case, for instance, of the biomedical field, where subcategorisation patterns should also include strongly selected modifiers such as location, manner and timing which are essential for the correct interpretation of texts (Tsai et al., 2007).

We have first developed our methodology for SCF acquisition on data extracted from an Italian general corpus. In a further step, we have applied the extraction process on an English biomedical corpus. This way, we have tried to evaluate both the effects of inter-linguistic variation, and

[1] The work reported in the paper has been carried out in the framework of the European BOOTStrep project (Bootstrapping Of Ontologies and Terminologies STrategic REsearch Project, FP6-028099), which aims at building lexical and conceptual repositories for the biology domain populated through text processing and mining from domain documents.

the biases deriving from the idiosyncrasies of a complex sub-language, such as the one typical of the biomedical domain.

2. Acquisition of Subcategorization Frames

Given a set V of verbs for which SCF information is to be acquired, automatic acquisition is performed through the following steps: syntactic annotation of the acquisition corpus; extraction of headword local contexts from the syntactically chunked texts; induction of potential subcategorisation frames. In what follows these steps are illustrated in some detail.

The starting point of our acquisition system is a syntactically shallow-parsed text corpus. In particular, we take as input a chunked text, that is a text that is segmented at the level of immediate non-recursive phrasal constituency which can be identified with certainty with no recourse to lexico-syntactic knowledge. Chunked syntactic representations are particularly suited for automatic lexical acquisition since the preliminary identification of syntactic chunks significantly reduces the search space for either arguments or modifiers (hereafter comprehensively referred to as "complements") of a verbal head in context.

For each verb v in V, all chunked contexts containing v are extracted from the syntactically pre-processed training corpora to form a set of "syntactically local contexts" of v (henceforth, SLCs).

The length of the extracted SLCs is reduced by applying a battery of linguistically-motivated constraints. This process, referred to as "context carving" (Federici et al. 1998), is achieved by scanning in SLC the sequence of chunks on the right and left side of v, to progressively include all adjacent chunks which are potentially dependent on v. Inclusion stops at a chunk in SLC which a battery of straightforward linguistically-motivated criteria considers as the likely initiator of a chunk sequence lying outside the dependency scope of v.

Through carving, SLCs are thus reduced to more reliable dependency islands, where noisy information is minimized. Carving is sensitive to a number of features ranging from chunk category to specific attributes in the chunk internal structure. To illustrate, consider the following SLC of the verb *chiudere* 'close down':

[N_C *lo yen*] [FV_C *ha chiuso*] [P_C *a Tokio*]
[P_C *a 120*] [I_C *dopo aver toccato*] [P_C *nel corso*] [P_C *della seduta*]
[N_C *il massimo*] [ADJ_C *storico*]
'the yen closed down in Tokyo at 120 after reaching the maximum ever in the course of the session'

Here carving leaves out chunks which are likely to depend on another headword than *chiudere*. In particular, the stop-chunk is identified here with the infinitival chunk (I_C): all chunks following it (underlined in the example above) are eventually left out. In more general terms, chunks are excluded from SLCs as potentially noisy information whenever the chunked context provides evidence for them to be understood as likely depending on a headword other than v. The newly Carved Contexts (CCs) are thus expected to contain adjacent chunks that are potential frame slots of v.

Extraction of SLCs and their carving represent the first steps towards subcategorization induction in the strict sense. SLCs represent a first level of abstraction from the real contexts in which verbal headwords occur, but obviously they cannot be considered potential frames in their own right yet. To assess their eligibility as possible frames we used a battery of discovery procedures which translate, in distributional terms, linguistic insights on the syntactic behaviour of verb complements as observed in the corpus.

In order to identify SCFs, a further set of linguistic heuristics is applied to CCs looking for the most likely frame slots (be they arguments or lexically-selected modifiers) of v. Starting from the assumption that all contextual chunks occurring immediately after the verbal headword are very likely governed by it, for each v in V the set of "potentially subcategorized slots" (PSS) has been identified. Selected PSS include frame slots occurring immediately after the verbal headword beyond a certain threshold (i.e. which cover at least 3% of the occurrences of v). Table 1 reports the typology of PSS as emerging from the corpus for the verbs *accettare* 'accept', *accusare* 'accuse' and *alludere* 'allude':

lemma	PSS
accettare	[CHE_C]
accettare	[I_C-*di*]
accettare	[N_C]
accusare	[I_C-*di*]
accusare	[N_C]
accusare	[P_C-*di*]
alludere	[P_C-*a*]

Table 1: Extracted PSSs for the verbs *accettare*, *accusare* and *alludere*.

The list of identified PSS is then used to drive the extraction process of potential subcategorization frames from CCs: a CC can be seen as a SCF instantiation if all its contextual chunks - or part of them - belong to the list of selected PSS.

Verb	SCF	p(SCF|v)
accettare	[N_C]	0.45
accettare	[]	0.33
accettare	[I_C-*di*]	0.13
accettare	[CHE_C]	0.05
accusare	[N_C]	0.38
accusare	[]	0.25
accusare	[N_C][I_C-*di*]	0.24
accusare	[N_C][P_C-*di*]	0.10
alludere	[P_C-*a*]	0.90

Table 2: Extracted SCFs for the verbs *accettare*, *accusare* and *alludere*.

SCFs are eventually reconstructed by grouping likely frame slots which happen to co-occur in the same CC

beyond a certain frequency threshold. We didn't consider the subject as part of SCF. Table 2 exemplifies the result of this acquisition step by reporting induced SCFs for the same selection of verbs, ordered by decreasing probability.

3. Experiments and evaluation of results

We tested our SCF acquisition algorithm on both Italian and English corpora. With regard to Italian, we worked on the chunked PAROLE Corpus (Goggi et al., 1997), a general corpus consisting of 3 million word tokens chunked with CHUG-IT (Federici et al., 1996). The English domain corpora (ca 6 million tokens) were provided by the European Bioinformatics Institute and chunked with the GENIA TAGGER v. 3.0 (Tsuruoka et al., 2005), a tagger specifically tuned for biomedical text[2]. In particular, we focussed on a list of 47 Italian communication verbs, and on 50 English verbs selected as particularly relevant for the biomedical domain.

Acquired SCFs can be evaluated in different ways, either through manual inspection, or against SCFs attested in existing lexicons, or by using them for different NLP tasks and applications. In this context, evaluation of extracted SCFs was carried out against gold standard resources in terms of type precision and type recall, measuring completeness and reliability of the bootstrapped syntactic lexicon: type precision was calculated as the percentage of correctly acquired SCFs with respect to all acquired SCFs, and type recall as the percentage of correctly acquired SCFs with respect to all SCFs attested in the gold standard lexicon. The f-measure was also computed to provide the weighted harmonic mean of precision and recall. F-measure was calculated as follows:

F = 2*(precision*recall)/(precision+recall)

If on the one hand this type of evaluation guarantees results which are comparable to those achieved through different techniques used across different languages, on the other hand it poses a considerable problem. In particular, the type recall measure does not always permit to discriminate between the effectiveness and reliability of the lexical acquisition system and the coverage of the reference lexical resource with respect to the acquisition corpus. In principle, to minimise this problem the reference lexical resource and the acquisition corpus should relate to the same domain but, as we shall see below, this may not always be the case.

In order to evaluate the effectiveness of our approach to SCF induction, we compared induced SCFs with the extracted carved contexts (CCs) which have been assumed as a baseline. In this way it is possible to evaluate the neat contribution of our SCF induction method.

3.1 Italian

Different gold standards were built from two Italian general resources.

1. IGS1: from a general purpose computational lexicon, the Italian PAROLE lexicon (Ruimy et al., 1997);
2. IGS2: from an Italian dictionary, Sabatini-Coletti (2006). This is actually the only Italian machine readable dictionary that describes verb subcategorization properties in terms of a limited number of valence frames;
3. IGS3: created by merging IGS1 and IGS2.

Table 3 contains the overall precision and recall for the selected set of communication verbs with respect to the different gold standards:

		IGS1	IGS2	IGS3
SCFs	Precision	42%	30%	52%
	Recall	8%	84%	78%
	F-measure	13%	44%	62%
baseline	Precision	23%	13%	27%
	Recall	72%	68%	65%
	F-measure	35%	22%	38%

Table 3: Evaluation of Italian results

It can be noted that the joining of the two gold standards (IGS1 and IGS2) causes an increase of the precision (52%) but a decrease at the level of recall (78%). In order to further validate acquired results, manual evaluation of acquired frames was also carried out, resulting in a much higher percentage of acquired correct frames, i.e. 93% (against the 40% of the baseline).

The significantly higher precision observed through manual inspection of acquired SCFs suggests that the reference lexical resource and the acquisition corpus are not well aligned, in the sense that there are frames which were correctly acquired but which were not recorded in the selected reference resources. This misalignment can also be seen as underlying the low recall which has been observed in all cases.

3.2 English

Different gold standards were built from different English reference resources.

1. EGS1: from a general purpose computational lexicon, the Valex5 Lexicon (Korhonen et al., 2006), which contains types and frequencies of filtered and smoothed verbal SCFs;
1. EGS2: from a general English dictionary, the Longman Dictionary (2006);
2. EGS3: from a biomedical English lexicon, the SPECIALIST Lexicon;[3]
3. EGS4: created by merging EGS1, EGS2 and EGS3.

Table 4 reports the overall precision and recall for the selected set of biologically relevant verbs. As it can be noticed, the merging of the three gold standards (EGS4) yields the best results, with an increase of precision (83%)

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/postagger/

[3] http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html

and a neglectable decrease of recall (51%) with respect to the domain-specific resource (EGS3). The comparison between the acquired SCFs and the SPECIALIST lexicon suggests that our system acquires many SCFs which are not considered as domain-relevant, but which are still correct (as they are included in the general reference resources): this explains the low precision score obtained with respect to the domain-specific resource.
By comparing acquired SCFs with the baseline, we record a significant improvement in precision and a slight decrease in recall which may be explained by the fact that SCF induction tends to discard low frequency SLCs, which are anyway recorded in the reference resources.

		EGS1+EGS2	EGS3	EGS4
SCFs	precision	69%	52%	83%
	recall	48%	54%	51%
	F-measure	57%	53%	63%
Baseline	precision	28%	17%	33%
	recall	52%	49%	53%
	F-measure	36%	25%	41%

Table 4: Evaluation of English results.

3.3 Comparative evaluation of Italian and English results

Both Italian and English results were evaluated against gold standard resources. Achieved f-scores are in line with state-of-the-art subcategorization acquisition systems (see Schulte im Walde, to appear), namely 62% in the case of Italian and 83% in the case of English. Concerning Italian, a wide range of variation is observed for what concerns the different gold standard resources; interestingly, the lowest f-score is obtained with respect to IGS1, a very rich general purpose computational lexicon which was not supposed to reflect corpus evidence. With regard to English, it is interesting to note that the lowest f-score was achieved with respect to the bio-medical English lexicon (EGS3). This suggests that an evaluation carried out against gold standard resources extracted from dictionaries is bound to the assumptions of the dictionary, which might differ from those in the lexical acquisition approach. A more flexible way to compare acquired SCF information is through manual inspection. This type of evaluation was carried out on acquired Italian SCFs only with encouraging results: we passed from 52% of precision to 93% of accuracy.

4. Verb clustering and the MDL Principle

Its merits notwithstanding, the methods we have presented in the sections above rely on the assumption that SCFs are lexically-specific properties of each verb. However, this runs against the robust evidence supporting the hypothesis that (at least part of) subcategorization properties depend on abstract semantic features of a verb. According to this view, verb lexicon is organized in terms of paradigmatic classes of verbs sharing similar semantic properties and similar SCFs. Since the seminal analysis of English verb semantic classes and syntactic alternations in Levin (1993), growing attention and efforts have been devoted to explore what is now referred to as the "syntax-semantics lexical interface", i.e. the correlations between verb semantic properties and type of selected syntactic frames. Interesting results on this topic have been carried out in computational linguistic analyses, such as for instance Merlo & Stevenson (2001), and more recently Schulte im Walde (2006). The latter work is particularly relevant to our research, because it compares an *a priori* semantic classification of German verbs with verb clusters automatically induced from the statistic distribution of verb co-occurrence with a pre-defined number of SCFs. The exploitation of semantic generalizations stemming form verb semantics may be extremely relevant for the unsupervised acquisition of SCFs. For instance, the fact that the verb *affermare* 'affirm' selects for a *che*–clause may be regarded not as an idiosyncratic feature of this verb, but directly derived from the fact that it is a verb of communication, and that members of this semantic class typically select for this kind of complement (e.g. *dire*, 'say', *credere* 'believe', etc.). Consequently, knowing that a verb belongs to a certain semantically motivated verb set could provide information to infer its syntactic selectional preferences with respect to a given SCF.
In the last part of our work, we intend to contribute to this issue by using the automatically extracted SCFs to test the hypothesis of the "syntax-semantics lexical interface". Starting form the SCFs extracted with the method illustrated in the former section, we want to induce clusters of verbs that share similar semantic properties. The issue is whether verbs sharing similar SCFs distributions happen to share similar semantic properties as well. To explore this question, we represent each verb with a n-dimensional vector reporting the verb statistical distribution with the automatically extracted SCFs. A clustering of verb vectors is performed using the Minimum Description Length Principle (MDL), a principle of data compression coming from Information Theory (Rissanen, 1989).
By finding the shortest model (or grammar) which best describes the data on hand, MDL gives an evaluation measure of the goodness of our analysis.
According to the MDL Principle, two lengths need to be calculated for each model, in order to determine the optimal one:

- length of the model in bits (model description length, L_m);
- length of the data described by the model, i. e. the cross-entropy of the corpus stochastically generated by the model and the original corpus (data description length, $L_{(D,m)}$).

On the one hand, the model length accounts for the set of linguistic units used by the model (and therefore it gives a picture of its complexity), on the other hand the data description length measures the accuracy of the model description (i. e. its fit to the data).

By exploiting the insight that "any regularity in the data can be used to compress the data, i.e. to describe it using

fewer symbols than needed to describe the data literally" (Grünwald, 2007), the MDL Principle states that the best model is the one minimizing the sum of these two lengths (total description length, L(m)):

$$M = \arg\min_m L(m).$$

4.1 State of the Art for MDL in linguistics

In the last decade MDL techniques have been applied to various tasks in automatic acquisition of linguistic structures.

Li & Abe (1998) introduce MDL in the context of case frame pattern acquisition. They start from an existing thesaurus and estimate the MDL-optimal tree cut model of a thesaurus tree for the given frequency data of a case slot. In this way they obtain case frame patterns whose fillers are optimally constrained by semantic restrictions defined over thesaural nodes, and use this information to resolve PP-attachment ambiguity. In a similar vein, McCarthy & Korhonen (1998) find alternating SCFs, that is frames where the same argument slots are syntactically realised in different structural positions, by comparing the model in which the frames are encoded separately with the one where corresponding arguments slots in the respective frames are combined. The model with the minimal description length is eventually chosen.

In a series of recent contributions, John Goldsmith (2001, 2006) uses MDL to proceduralize Harrisian discovery procedures for morphological segmentation. Starting form the assumption that morphological information about a language can hardly be reduced to *local* information about letter bigrams or trigrams of that language, Goldsmith frames the task as a data compression problem: "find the battery of inflectional markers forming the shortest grammar that best fits training evidence", where i) a grammar is a set of *paradigms* (named *signatures*) defined as lists of inflectional markers applying to specific verb classes and ii) the training evidence is a text corpus. The task is a top-down global optimization problem and boils down to a grammar evaluation procedure. Given a set of candidate inflectional markers, their probability distribution in a corpus and their partitioning into paradigms, MDL allows calculation of i) the length of the grammar (in terms of number and size of its paradigms) and ii) the length of the corpus generated by the grammar (*i.e.* the set of inflected forms licensed by the grammar according to a specific probability distribution). In MDL, the notion of length is derivative of the information theoretic notion of the number of bits required to encode linguistic units, whether they are stems, suffixes or word tokens. Intuitively, minimising the length of the corpus in bits requires that very frequent tokens should be assigned a shorter bit code than less frequent tokens. Minimising the length of the grammar, on the other hand, requires that frequently used paradigms are given preference to rarely used ones, as the cost of encoding a rare paradigm in bits is very high. Hence, a good language model is the one where the sum of the length of the grammar and the length of the corpus generated according to the probability assigned by the grammar is smallest. This policy disfavours two descriptively undesirable extremes: a corpus-photograph model, with a very long grammar where each verb form has, as it were, a paradigm of its own, such that the inflected forms generated by the grammar have the same probability distribution found in the corpus; and a very short but profligate model, with one paradigm only, where any verb combines with any marker according to the product of their independent probability distributions, thus generating many word forms that are not attested in the training corpus (including *goed* for *went*, *stricked* for *struck*, *bes* for *is* etc.).

4.2 Clustering of verbs using the MDL Principle

We adapted Goldsmith's grammar evaluation procedure to our task. In our case, the dataset consists of all couples <verb, SCF> acquired from the corpus, together with their frequency distributions. The candidate models to be evaluated define possible groupings of verbs according to similarity in the distribution of their frames. By clustering verbs that share frames with the same distribution, one gains on model length as more SCFs are associated with a single verb class instead of being independently stipulated for each individual verb. However, something is lost on modelling data distributions, as the original distributions are averaged out over all clustered verbs. [4] An additional source of length in the model is that identification of a new verb cluster requires introduction of a new data structure, thus adding an extra cost to the overall model length. More concisely, the overall length of a model M is given by the following equation:

$$L(M) = L(\text{frames}) + L(\text{verb classes}) + L_M(\text{corpus}).$$

MDL is thus used to estimate exactly such a trade-off between the length of the model and the length of the corpus distribution generated by the model in such a way as to minimize L(M). This is done iteratively, according to the following steps:

1. Let $\{v_1,\ldots,v_r\}$ be the verbs in question. At the first step, a baseline model M_0 is considered where each verb belongs to a different class $\sigma_j=\{v_j\}$, j=1,..,r. This model offers the best data fit, but it is also the lengthiest one, as all <verb, SCF> patterns are listed independently. At each i^{th} clustering step, the baseline model is the model M_{i-1} obtained at the previous clustering step.
2. At the second step, M_0 is compared with any model $M_1(h,k)$ (h,k=1,..,r. and h≠k) consisting of the new class $\sigma_{r+1}=\{\sigma_h, \sigma_k\}$ plus all other classes $\sigma_j=\{v_j\}$ (j=1,..,r and j≠h,k.). By doing this way, we try to build a more compact model where $\{v_h\}$ and $\{v_k\}$ share the same frames (thus decreasing the model description length). At the same time, however, we increase the number of linguistic units making up the model by introducing one more verb class σ_{r+1} and worsen the model's fit to

[4] This criterion allows for two classes to be merged even if either of them presents a frame which is not attested with the other, as long as the difference in probability is not too big and allows a decrease in the total description length.

the original data distribution. We then look for the optimal balance between these two tendencies, by selecting only those clusters where *difference*(0,1(h,k)) = $L(M_0) - L(M_1(h,k)) \geq 0$, and choosing the couple of indices $\{n_1,m_1\}$ such that

$$M_1(n_1,m_1) = \arg\max_{(h,k)}(\textit{difference}(0,1(h,k))).$$

This means both that the model M_1 is compact enough to increase the conciseness of the analysis in spite of the loss in data fit, and that this increase is the highest.

3. At each iteration, only the class pair minimizing the sum of model length is allowed to form a new verb class. At step i+1, the baseline $M_i(n_i,m_i)$ is compared with all the models $M_{i+1}(h,k)$ (for $h,k=1,..,r+1$, $h \neq k$ and $h,k \neq n_i,m_i$) and the couple of indices $\{n_{i+1},m_{i+1}\}$ is chosen only if

$$M_{i+1}(n_{i+1},m_{i+1}) = \arg\max_{(h,k)}(\textit{difference}(i(n_i,m_i),i+1(h,k))),$$

provided that *difference* is positive.

4. We keep clustering this way until *difference* stops increasing, acquiring the final set of verb classes and the frames associated with them.

As a result of this process, verbs are iteratively grouped into a (sometimes incomplete) binary branching hierarchy. When applied to the 47 Italian communication verbs, this method stopped after 23 clustering steps, generating the clusters shown in Figure 1. The progress of the total length of the models from the baseline model to the last clustering step is displayed in Figure 2.

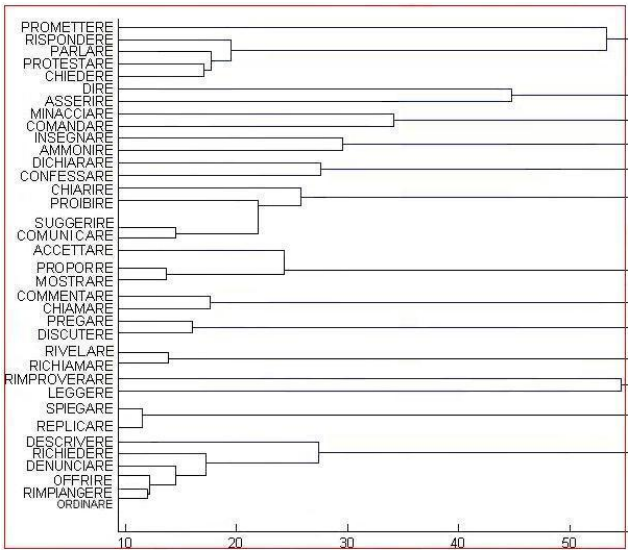

Figure 1: The results of the MDL-clustering for the 47 Italian communication verbs.

The verb classes obtained this way are assigned a new cluster-based frame distribution, as exemplified in Table 5 for 4 clustered Italian communication verbs where the first four rows represent individual verbs and part of associated frame probabilities, and the last one the MDL-based verb cluster with probabilities associated to the frames for the cluster. It is interesting to note that at the verb cluster level it is possible to generalise over the evidence attested in the corpus for individual verbs. For instance, among the frames associated with the cluster there is also the [P_C-*a*] frame which emerged only from the occurrences of the verbs *comunicare* 'communicate' and *suggerire* 'suggest' but which appears to be a valid frame also for the other verbs in the cluster. A preliminary qualitative analysis of induced verb clusters shows that MDL-based clustering represents a promising line of research which is worth being pursued to go beyond evidence attested in the corpus for individual verbs and to explore the syntax-semantics lexical interface. No quantitative evaluation of this clustering step has been carried out yet.

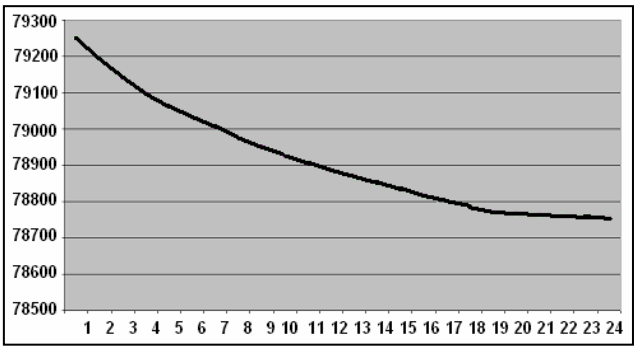

Figure 2: Total lengths of the models corresponding to the 23 clustering steps for the Italian communication verbs.

	[]	[*che*_C]	[1_C-*di*]	[N_C]	[P_C-*a*]	[*perché*_C]
chiarire 'clarify'	0.34	0.10	0	0.40	0	0.009
comunicare 'communicate'	0.24	0.15	0	0.31	0.08	0
proibire 'forbid'	0.21	0.03	0.03	0.51	0	0
suggerire 'suggest'	0.24	0.10	0.009	0.42	0.02	0.02
verb class (cluster)	0.25	0.10	0.008	0.41	0.02	0.02

Table 5: The probabilities associated to the frames for the Italian verbs *chiarire*, *comunicare*, *proibire*, *suggerire* and for their class, after MDL-clustering.

5. Conclusions

In this paper, we presented a discovery approach to SCF acquisition. The proposed technique operates on syntactically shallow-parsed corpora on the basis of a limited number of search heuristics that make no reliance on built-in lexico-syntactic knowledge about SCFs. Experiments have been carried out on Italian and English, and on different text types, i.e. general corpora in the case of Italian and biomedical texts in the case of English. Although preliminary, results are in line with state-of-the-art lexical acquisition systems.

Starting from the assumption that subcategorization properties should not be seen as idiosyncratic properties of individual verbs but rather as depending on abstract semantic features, we used acquired SCFs to test the hypothesis of the "syntax-semantics lexical interface". In particular, the issue of whether verbs sharing similar SCFs distributions happen to share similar semantic properties as well was explored by clustering verb vectors using the

Minimum Description Length Principle (MDL). First experiments in this direction were carried out on Italian verbs with encouraging results. In the near future, we expect to evaluate these results more extensively, with respect to both coherence of the obtained lexico-semantic clusters and coverage of the subcategorization behaviour of clustered verbs. Since frequency distributions are recalculated relative to MDL-optimal verb classes (rather than individual verbs), we expect rare SCFs and rare verbs to be better represented in the clustering model than in the original training data.

6. Acknowledgments

The work described in this paper has been funded by the European BOOTStrep project (FP6 - 028099).

7. References

Alonso, L. et al. (2007). Obtaining coarse-grained classes of subcategorization patterns for Spanish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP'07*. Borovets, Bulgaria, pp. 30--34.

Basili, R., Pazienza, M.T., Vindigni, M. (1997). Corpus-driven Unsupervised Learning of Verb Subcategorization Frames. In *Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence on Advances in Artificial Intelligence*. London, UK : Springer-Verlag, pp. 159--170.

Bourigault, D. & Frérot, C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan, France.

Brent, M. R. & Berwick, R. C. (1991), Automatic acquisition of subcategorization frames from tagged text. In *Proceedings of the Workshop on Speech and Natural Language, Human Language Technology Conference*. Pacific Grove, California, pp.342--345.

Briscoe, T. & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied natural language processing*. Washington, DC, pp.356--363.

Federici S., Montemagni S., Pirrelli V. (1996). Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In *Proceedings of the Workshop On Robust Parsing, in the framework of ESSLLI-96*. Prague, 12-16 August 1996.

Federici, S., Montemagni S., Pirrelli V., Calzolari N. (1998). Analogy-based Extraction of Lexical Knowledge from Corpora: The SPARKLE Experience. In A. Rubio, N. Gallardo, R. Castro, A. Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain. Vol. I. 75-82.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2), pp. 153--198.

Grünwald, P. (2007). *The Minimum Description Length Principle*. The MIT Press.

Korhonen, A. (2002), Subcategorization acquisition, University of Cambridge. Ph.D. thesis. University of Cambridge.

Korhonen, A. et al. (2006). A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genova, Italy.

Goggi, S. et al. (1997). Italian Corpus Documentation, LE-PAROLE WP2.11. ILC, Pisa.

Levin, B. (1993), *English verb classes and alternations: a preliminary investigation* Chicago: The University of Chicago Press.

Li, H., Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24, pp. 217-244.

McCarthy, D., Korhonen, A. Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*. Montreal, 2, pp. 1493--1495.

Merlo, P., Stevenson, S. (2001), "Automatic Verb Classification based on Statistical Distributions of Argument Structure", *Computational Linguistics*, 27(3): 373-408.

Rissanen, J. (1989), *Stochastic complexity in statistical inquiry* Singapore: World Scientific Publishing Co.

Ruimy, N. et al. (1998). LE PAROLE Project: the Italian Syntactic Lexicon. In *Proceedings of Euralex-98*. Liège, France.

Schulte im Walde, S. (2006), "Experiments on the Automatic Induction of German Semantic Verb Classes", *Computational Linguistics*, 32(2): 159-194.

Schulte im Walde, S. (to appear). The Induction of Verb Frames and Verb Classes from Corpora. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

Tsai R.T.H. et al. (2007). BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, 8: 325.

Tsuruoka, Y., Tateishi, Y., Kim, J-D., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382-392.

Ushioda et al. (1993), The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In B. Boguraev & J. Pustejovsky (Eds.), *SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text.*. Columbus, Ohio, pp. 95--106.

Utsuro, T. et al. (1997), Maximum Entropy Model Learning of Subcategorization Preference. In *Proceedings of the 5th Workshop on Very Large Corpora*. Beijing, China, pp. 246--260.

Daniel Zeman and Anoop Sarkar (2000), Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*. Athens, Greece, pp. 227-233.