Carving Verb Classes from Corpora

Alessandro Lenci

Università di Pisa, Dipartimento di Linguistica "T. Bolelli"
alessandro.lenci@ling.unipi.it

Abstract

In this paper, we discuss some methodological problems arising from the use of corpus data for semantic verb classification. In particular, we present a computational framework to describe the distributional properties of Italian verbs using linguistic data automatically extracted from a large corpus. This information is used to build a distribution-based classification of a set of Italian verbs. Its small scale notwithstanding, this case study will provide evidence for the complex interplay between syntactic and semantic verb features.

1. *Classifying verb meanings*[*]

Understanding how verbs can be classified according to their common semantic features is a major goal shared by lexical semanticists, computational linguists and cognitive scientists alike. In fact, important generalizations about a verb behavior can be stated by referring to its semantic class. However, the term that best describes the current research on verb classes is "embarrassment of richness". Several semantic verb classifications are already available for English: *WordNet* (Fellbaum 1998), *VerbNet* (Kipper-Schuler 2005), *FrameNet* (Fillmore et al. 2003), *Levin Classes* (Levin 1993), just to cite the most prominent examples. In Italian, we have the WordNet-style semantic classifications provided by *Italian WordNet* (Pianta et al. 2002) and *ItalWordNet* (Roventini et al. 2000) – each based on very different criteria - and the system of verb classes in *Simple* (Lenci et al. 2000), which is partly inspired to the Generative Lexicon (Pustejovsky 1995). As Čulo et al. (2008) point out, the main shortcoming of this situation is that existing semantic verb classifications may vary dramatically, not only with respect to their granularity (i.e. the number of semantic classes), but also with respect to the criteria with which the class system is organized, thereby often resulting in different, even orthogonal classifications of verbs.[1]

One reason explaining this wide spectrum of variation is the fact that there are two main approaches to semantic verb classification, which we will refer to as *ontology-based* and *distribution-based*. Their main difference between them lies in the extent to which the distributional properties of verbs, i.e. the set of linguistic constructions and patterns they occur with, is adopted as the main criterion for class identification and class membership. In ontology-based classification

schemes, such a criterion is provided by the features of the extra-linguistic event or situation expressed by a verb meaning, rather than by its linguistic behavior. One clear example of this strategy is FrameNet. In fact, in FrameNet two verbs belong to the same semantic class to the extent they evoke the same frame, which is interpreted as a conceptual, schematic representation of a situation. For instance, the verbs *eat* and *devour* are grouped together because they evoke the same Ingestion Frame, corresponding to the schematic representation of a situation in which "an Ingestor consumes food or drink (Ingestibles), which entails putting the Ingestibles in the mouth for delivery to the digestive system".[2] Conversely, Levin Classes can be regarded as an example of distribution-based classification, because the main membership criterion is the range of syntactic alternations licensed by a verb, as a key aspect of its syntagmatic, distributional properties. Therefore, since *eat* but not *devour* allows for object drop and conative constructions, in Levin Classes these verbs do not belong to the same class (Levin 1993: 213-215). As expected, most verb classifications tend to mix both criteria, thereby resulting in a huge spectrum of alternative possibilities depending on the salience of the ontological or distributional perspective in designing the semantic classes.

Since the ontology-based and the distribution-based criteria often produce orthogonal results, we need to find arguments that help to decide between them and to identify the most appropriate methodology for semantic verb classification. Here, we will present four reasons supporting the claim that *distributional data should be regarded as the main (possibly the only) constraint for semantic class design*:

1. distributional data *de facto* represent the most robust "*observables*" that are available to us to reconstruct verb meaning and to define the proper membership criteria of semantic equivalence classes. It is instead highly risky, and even sometimes unwarranted, to ground a classification scheme on our intuitions or presumptions about the unfolding of extra-linguistic situations and events. The hypothesis that members of the same semantic class refer to events sharing a certain number of features is surely appealing, but still it raises the problem of finding effective, independently motivated, objective criteria to establish the conceptual features relevant to estimate verb semantic similarities. At least part of the large degree in variability in semantic verb classifications is indeed due to the lack of such precise identity criteria of the relevant semantic features grounding the class choice. While distribution-based classifications are supported by linguistic data, a similar set of observable data is not available to date to support ontology-based generalizations;

2. if we adopt a strict distributional perspective, verb semantic classes can be regarded as a kind of "latent variables" that are responsible for the distributions of the linguistic constructions we observe, and that we try to uncover by inspecting a significant amount of such empirical

distributions. In fact, wide empirical evidence supports the existence of a close *correlation* between semantic content and the way verbs are used in linguistic contexts and constructions. Levin's proposal to ground verb semantic classifications on the verb diathesis alternations can indeed be regarded as a particular instance of the so-called *Distributional Hypothesis* (DH; Harris 1954, Miller & Charles 1991, Lenci 2008). According to the DH, at least certain aspects of the meaning of lexical expressions depend on their distributional properties in the linguistic contexts, and the degree of semantic similarity between two linguistic expressions *A* and *B* is a function of the similarity of the linguistic contexts in which *A* and *B* can appear. Therefore, it is at least possible to exploit such correlations to identify the groups of verbs that pattern alike, searching for the elements of meaning they share. Distributional data can be used as "a probe into the elements entering into the lexical representations of word meaning" (Levin 1993:14);

3. the distribution-based approach seems to be more suitable if we are interested in classifying the meaning of verbs, *qua* linguistic objects. This fact is often overlooked in the linguistic and cognitive literature, in which an equation between meanings (as linguistic entities) and concepts (as mental - not necessarily linguistic - representations of categories of extra-linguistic entities) is assumed. Here, we share the position defended by Vigliocco & Vinson (2007), who argue that such an equivalence can not be presupposed. Conceptual representations and semantic contents should rather be conceived as distinct, although deeply interconnected layers. Therefore, if we are interested in understanding the meaning of *eat*, and in identifying the verbs that are semantically similar to *eat* and are to be grouped together in the same class, we should rather consider how these verbs are used in linguistic constructions, rather than looking at the way eating events occur. Once we have properly designed our distribution-based verb semantic scheme, we can use it to study the way events are conceptually represented and expressed, thereby avoiding (or reducing) the risk of circularity, since the classification scheme would now be independent from any unwarranted assumption about conceptual structures;

4. the current availability of large-scale corpora, tools for natural language processing and automatic text annotation, and statistical methods to extract linguistic data from texts allow us to turn the distributional method into a powerful and really effective criterion for exploring verb behavior. This does not entail that the distributional approach should only be corpus-based. The distribution-based method for verb semantic classification – at least as is conceived here – consists in assuming as the main criterion for verb semantic similarity and semantic class design the way verbs occur with linguistic constructions. Therefore, there is nothing in principle that prevents this method to be carried out by using corpus data along with carefully elicited speakers' judgments about the distributional properties of lexical items. However, speakers' intuitions are often not

reliable and are not susceptible to frequency-based analyses. Conversely, computational linguistics methods applied to large-scale corpora provide us with an extremely rich array of evidence about lexical distributions. This is the reason why corpus data the preferred evidence for the distributional approach to verb class construction. Moreover, corpus derived statistics can be used to estimate the salience of verb constructions and to characterize verb properties as continuous variables. Thus, classifications can be sensitive not only to the constructions a verb can occur with, but also to their different degrees of statistical salience.

In this paper, we will discuss some methodological problems arising from the use of corpus data to bootstrap verb semantic classes. There are indeed two main issues concerning distribution-based verb classifications: 1) *which type of information can be extracted from corpora to characterize a verb distributional behavior?* 2) *how to use this information to partition verbs into semantic equivalence classes?* State-of-the-art methods in computational linguistics provide answers to both questions, but, as we will show below, with different degrees of success. We will address the first issue by presenting a computational framework to describe the distributional properties of Italian verbs using linguistic data automatically extracted from a large corpus (Section 2). Then, this information will be used to build a distribution-based classification of a set of Italian verbs (Section 3). Its small scale notwithstanding, this case study will show the need to model the complex interplay between syntactic and semantic verb features as a precondition to meet the challenge of verb classification.

2. ***Profiling verb distributions***

The *distributional profile* of a verb *v* is defined here as an array of information extracted from a corpus to characterize the distributional properties of *v*. The automatic acquisition of verb information from corpora represents a longstanding research avenue in computational linguistics (Manning & Schütze 1999). Efforts have mostly focused on developing methods to extract verb subcategorization frames (Schulte im Walde 2009), to identify verb selectional preferences (Light & Greiff 2002), and (though to a less extent, given the challenging nature of the task) to automatically detect diathesis alternations (McCarthy 2001). In this section, we will describe the application of some of these computational methods to extract distributional profiles of Italian verbs from *La Repubblica* (Baroni et al. 2004), a corpus of ca. 326 million word tokens of newspaper texts. The corpus was first lemmatized and part-of-speech tagged, and then parsed with DeSR (Dependency Shift-Reduce), a state-of-the-art stochastic dependency parser (Attardi & Dell'Orletta

2009). For the 3,873 most frequent verbs (min. freq. = 100; max. freq. = 830,903), their distributional profile has been extracted from the parsed corpus. Each profile is in turn organized into a *syntactic profile (*Section 2.1) and a *semantic profile* (Section 2.2), respectively encoding the syntagmatic and semantic distributional properties of the verb.[3]

2.1. *Syntactic profiles*

The syntactic profile of a verb *v* a list of its subcategorization frames (SCFs), ordered by their statistical salience for the verb. Each SCF corresponds to a specific pattern of syntactic dependencies headed by *v*. It is formed by an unordered set of *slots* (e.g., subject + complement introduced by the preposition *a* + direct object) and is identified by a synthetic label (e.g., SUBJ#OBJ#COMP-*A*). Among the subcategorizing elements we also considered the reflexive pronoun *si*. The zero-argument construction (labeled with SUBJ#0) instead represents the case in which the verb appears with no dependencies, besides the (optional) subject. For instance, the sentences *Gianni ha pianto* "John cried" and *Il vaso si è rotto* "The vase *si*-broke" are respectively instances of the frames SUBJ#0 and SUBJ#SI#0. We did not formally encode in the SCFs the distinction between arguments and adjuncts. Therefore, the sentences *Gianni abitava in città* "John lived in town" and *Gianni mangiava in città* "John ate in town" are regarded as instances of the same SCF SUBJ#COMP-*A*. This is essentially due to the limitations of the DeSR parser, which abstracts away from this distinction (like most state-of-the-art parsers do). In fact, arguments and adjuncts are notoriously hard to discriminate, let alone for natural language processing systems. We leave to future research how to capture this contrast with automatically derived distributional data.[4]

The process of syntactic profiling was carried out in the following way:

- we hand-selected 100 SCFs among the most frequent syntactic dependency combinations in the corpus (abstracting from linear order; i.e. *Gianni ha dato il libro a Maria* "John gave the book to Mary" is considered to be the same dependency pattern as *Gianni ha dato a Maria il libro* "John gave the book to Mary");
- for each selected verb *v*, we computed its joint frequency with each SCF, based on the verb dependency patterns automatically extracted form the parsed corpus;
- verb-SCF frequency was then used to compute the Local Mutual Information (LMI) score (Evert 2008), to estimate the statistical salience of the SCF for *v*. LMI is a variant of the Pointwise Mutual Information, to avoid its bias towards overestimating the significance of low frequency events. This score is normally used for the study of lexical collocations, and was applied here to identify the most prototypical SCFs of a verb.

Table 1 reports a sample syntactic profile extracted for the verb *arrivare* "arrive". The association score highlights the most prominent SCFs for this verb, e.g. the prepositional complement headed by *a* (cf. *Gianni è arrivato a casa* "John arrived at home), the infinitival clause introduced by the same preposition (cf. *L'acqua è arrivata a lambire la strada* "The water arrived at touching the road"), etc.

@@ Insert Table 1 here

2.2. *Semantic profiles*

A key aspect of the distributional properties of a verb is represented by the semantic type of the lexemes realizing its syntactic slots, i.e. its *slot fillers*. For instance, both *uccidere* "kill" and *mangiare* "eat" occur with the transitive SCF SUBJ#OBJ, but the former typically selects for animate direct objects, while the latter typically selects for foods. Characterizing the verb combinatorial semantic constraints, i.e. its selectional preferences, is notoriously a hard task. Adopting a distributional perspective, the selectional preferences of a verb can be obtained through an inductive generalization from the prototypical lexical fillers of the verb syntactic slots. This is again an instance of the DH illustrated in Section 1. In fact, the similarity between two verbs with respect to the semantic constraints in a given syntactic position (e.g., the direct object) can be regarded as a function of the similarity of the lexical items that can occur in that position (Erk 2007).

Consistently with the distributional approach, the semantic profiles extracted for the Italian verbs are two-layer structures specifying the following information for each SCF f_v of a verb v and for each slot s of f_v:

i) the *lexical set* of s (Hanks 1996, Hanks & Pustejovsky 2005), defined as a set of the fillers of s, ranked by their degree of prototypicality. For example, the lexical set of the direct object slot of the verb *leggere* "read" is composed by *libro* "book", *giornale* "newspaper", *rivista* "magazine", and so forth;

ii) (only for noun-selecting slots) the *selectional preferences* of s, defined as a ranked list of the noun semantic classes (e.g. PERSON, ANIMAL, etc.) that best describe the semantic types of the fillers of s, i.e. the semantic constraints of s. Currently, the selectional preferences have been characterized in terms 24 broad semantic classes, corresponding to the so-called "top nodes" dominating the semantic noun taxonomy in the Italian section of MultiWordNet (Pianta et al. 2002): ANIMAL, ARTIFACT, ACT, ATTRIBUTE, FOOD, COMMUNICATION, KNOWLEDGE, BODY_PART, EVENT, NATURAL_PHENOMENON,

SHAPE, GROUP, LOCATION, MOTIVATION, NATURAL_OBJECT, PERSON, PLANT, POSSESSION, PROCESS, QUANTITY, FEELING, SUBSTANCE, STATE, TIME. [5]

The frequency of a lexeme occurring in a slot *s* was used to estimate with LMI its prototypicality as a filler of that slot. Then, the lexical set of *s* was obtained by selecting only the fillers with LMI greater than 0. In turn, nominal lexical sets were used to compute the selectional preferences with the following variation of the algorithm described in Schulte im Walde (2006):

- the co-occurrence frequency of each noun filler of a verb slot *s* was uniformly divided among the different senses assigned to the noun in MultiWordNet;
- the sense frequency was then propagated up to the semantic hierarchy to the 24 mutually exclusive top-nodes, thereby obtaining the joint frequency between *s* and each of the WordNet top-classes.
- as an element of novelty with respect to Schulte im Walde (2006), we calculated the LMI association score between each *s* and each semantic class. The semantic classes with LMI greater than 0 were then selected to represent the selectional preferences of *s*.

Table 2 reports a complete semantic profile for the SCF SUBJ#OBJ#COMP-*A* of *comunicare* "communicate", with the top part of the lexical sets associated to each frame slot and the semantic classes that describe their selectional preferences:

@@ Insert Table 2 here

Distributional semantic profiles have both a descriptive and a predictive function. On the one hand, lexical sets provide a sort of "snapshot" of the nouns occurring in a corpus with a verb in a certain syntactic position, together with an estimation of their statistical salience. On the other hand, selectional preferences represent a way to generalize from these instances to more abstract semantic properties of the verb arguments, thereby making predictions about previously unseen slot fillers. This information is also useful to compare verbs with respect to their semantic combinatorial constraints. For instance, Table 3 reports the verbs in the corpus with the highest association scores respectively to the class PERSON and LOCATION as the preferred semantic type selected by the prepositional complement introduced by *a* in the SUBJ#OBJ#COMP-*A* frame:

@@ Insert Table 3 here

Looking at Table 3, we can notice that, despite their *prima facie* similarity, the verb *mandare* "send" radically differs from the verbs *consegnare* "deliver" and *inviare* "send", as for the type of

the semantic constraints on the COMP-*A* slot. This is also confirmed by the whole spectrum of semantic classes associated with this slot (cf. Table 4).

@@ Insert Table 4 here

Even if the three verbs can be used almost interchangeably in some contexts, these data reveal a strong distributional contrast pointing towards major differences in their semantics. From the fact that *consegnare* "deliver" prefers human-like, animate oblique arguments, we can infer that delivering implies that there is someone who is able to receive what is delivered. On the other hand, *mandare* "send" does not have such an entailment, and can simply express a displacement of an object to another location. Moreover, the near-synonym *inviare* actually differs from *mandare* because it expresses events whose typical oblique arguments are animate (e.g., persons, institutions, etc.), like *consegnare*.

Distributional profiles provide us with a very large array of corpus-based information about the syntagmatic and semantic constraints to which verbs obey. The profiles built for the Italian verbs include the SCFs with which the verbs co-occur, together with the slot fillers and semantic types selected by these SCFs. Moreover, simple statistical association scores give an estimation of the relative degree of prototypicality of the different bits of information in the profile. It is also worth remarking that the information concerning the semantic classes selected by verbs is also fully distribution-based. Even if we have assumed a background semantic classification for the nouns, i.e. the semantic hierarchy provided by WordNet and the list of its top-nodes, still the association between a verb and the classes it selects for is totally data-driven, and grounded on the statistical distribution of its noun fillers.[6]

3. From distributional profiles to semantic classes

Computational linguistics research has produced an increasingly large number of methods for the automatic induction of verb classes from corpus data (cf. for instance Merlo & Stevenson 2001, Lapata & Brew 2004, Schulte im Walde 2006, Joanis et al. 2008, Li & Brew 2008, Sun & Korhonen 2009, among many others). Behind the differences, it is possible to identify a common approach to the problem of verb classification. First of all, verbs are represented as numerical vectors, whose dimensions correspond to a statistical weight derived from the verb joint frequency with a certain number of distributional features extracted from corpora with methods similar to

those illustrated in Section 2. Computational models differ for the type of distributional features adopted, such as the lexical collocates of a verb, SCFs, SFCs enriched with slot fillers and/or selectional preferences, or some combination thereof. Secondly, verb classification is usually modeled either as a supervised classification task (Merlo & Stevenson 2001, Joanis et al. 2008), or as an unsupervised clustering task (Schulte im Walde 2006, Sun & Korhonen 2009), using verb vectors as input. Again, a large spectrum of variation arises from the choice of the particular clustering or classification algorithm.

The state of the art in computational methods for automatic verb induction has achieved promising results, which shed light on the predictive power of different types of distributional features for verb classification (cf. Korhonen 2009 for a survey). However, the common goal of most of these approaches is to find reliable automatic methods to classify verbs against a given class repertoire, rather than to "discover" verb classes. Indeed, some form of verb semantic classification is presupposed by all existing methods, whose standard approach is to choose a sample of verbs, run a clustering or classification algorithm and evaluate the results against a "gold standard" semantic classification. Efforts are focused on identifying the feature combination and/or classification algorithm that best approximates the *a priori* classification. Most current work has in fact been carried out on English, using Levin Classes as background classification. Its advantages notwithstanding, this is not a suitable approach for languages, such as Italian, still lacking a wide-coverage, Levin-style verb classification. It is also worth noticing that even for English few attempts at extending Levin Classes with corpus data have been carried out in computational linguistics. For instance, Kipper-Schuler et al. (2008) have extended Levin Classes to cover verbs with sentential complements (not included in the original classification), but the new classification has been carried out manually, using distributional features (i.e. SCFs) extracted automatically from a corpus.

There is also a theoretical reason that makes automatic verb classification still unreliable. Most current methods use hard clustering algorithms, which assign verbs to one class only, thereby being essentially unable to cope with verb polysemy, and the consequent need for multiple class assignments. Moreover, each verb is represented just by one vector recording its global distributional history, i.e. all the different contexts in which has been observed in the corpus. The major shortcoming of this approach is that different usages of a verb end up being squeezed on the same vector. Consequently, it is impossible to capture the fact that alternative distributional patterns of a verb may be linked to different meanings and point to different verb classes. Therefore, there is a serious risk of oversimplifying the complex interaction between syntactic distributions and the semantic features that are relevant for verb classification.

The general conclusion to be drawn is that automatic methods are still substantially unreliable to induce a distribution-based verb classification. Semi-automatic approaches similar to the one in Kipper-Schuler et al. (2008) are instead more promising: *verb distributional profiles are first automatically extracted from large corpora and then distributionally coherent verb classes are carved from these profiles*. In the following section, this method will be illustrated in a small-scale case study of Italian verb classification.

3.1. *A case study in distribution-based classification of Italian verbs*

We are going to present a simple method to build a distribution-based classification of Italian verbs semi-automatically, consisting of the following steps:

- first a specific distributional pattern, in the present case a SCF, is chosen as a "seed" for verb selection and classification;
- then, the verbs in the corpus that have this SCF in their syntactic distributional profile are identified;
- finally, the selected verbs are partitioned into classes taking into account their distributional profile, i.e. the other SCFs and selectional preferences.

For the purpose of this paper, we have chosen as "seed pattern" the SCF SUBJ#OBJ#INF-*A*, corresponding to a frame formed by a subject, a direct object and an infinitival clause introduced by the preposition *a*:

(1) [$_{\text{SUBJ}}$ *Gianni*] *ha costretto* [$_{\text{OBJ}}$ *Maria*] [$_{\text{INF-}A}$ *a partire*].
"John forced Mary to leave"

This is a specific and fairly complex construction, which offers an interesting vantage point on the interaction between syntactic patterns and meaning dimensions. The Italian verbs extracted from *La Repubblica Corpus* that have this SCF as part of their distributional profile are reported in the Appendix. They have been grouped into classes according to similarities in their syntactic and semantic distributional profiles. Each class has also been annotated with its most distinctive distributional features, that represent a sort of "distributional signature" for the class. The purpose of this section is to discuss the criteria behind this proposed classification.

The first thing to notice is that there is a small group of verbs that can clearly be set apart from the rest of the verb sample. These verbs form a semantically homogenous class, the *Trascorrere* verbs, whose members occur with the SCF SUBJ#OBJ#INF-*A*, with the OBJ slot selecting

for nouns referring to temporal entities or events, e.g. *Gianni ha trascorso la giornata / partita a leggere* "John spent the day / the game reading". The infinitival clause denotes an event performed by the verb subject during the time or situation expressed by the direct object.

The rest of the sample include verbs whose OBJ slot in the SCF SUBJ#OBJ#INF-*A* is filled by nouns of semantic type PERSON or GROUP, i.e. referring to human or human-like entities (e.g., institutions). This set can in turn be carved into various classes, once we consider the other distributional patterns of the verbs. For instance, in the *Scoraggiare* class the SCF SUBJ#OBJ#INF-*A* can be alternatively realized as a SCF with a direct object and a nominal infinitive headed by *da*, as shown by these examples from *La Repubblica* corpus:[7]

(2) a. *Noi abbiamo sconsigliato* [$_{OBJ}$ *Andreotti*] [$_{INF\text{-}A}$ *a proseguire*].
"We did not advise Andreotti to go on"
b. *I leader della DC hanno sconsigliato* [$_{OBJ}$ *Andreotti*] [$_{INF\text{-}DA}$ *dall' insistere sul decreto*].
"The DC leaders did not advise Andreotti to insist on the decree"

The verbs in the *Autorizzare* class instead alternate the SCF SUBJ#OBJ#INF-*A* with a SCF with a direct object:

(3) a. *Gianni ha sollecitato* [$_{OBJ}$ *Maria*] [$_{INF\text{-}A}$ *a partire*].
"John urged Mary to (*a*) leave"
b. *Gianni ha sollecitato* [$_{OBJ}$ *la partenza di Maria*].
"John urged Mary's departure"

Notice that there is also a meaning shift between the two variants, with only (3a) entailing that Mary was "directly" urged to leave by John.

The *Consigliare* and *Convincere* classes can instead be distinguished by other frames that contribute to shape their "distributional signature". For instance, with the *Consigliare* verbs the SUBJ#OBJ#INF-*A* (4a) can be found along with the frame SUBJ#OBJ#INF-*DI* (4b), and with the frame SUBJ#COMP-*A*#INF-*DI* (4c), as shown by these examples from *La Repubblica*:

(4) a. *La paura per il terrorismo sta infatti consigliando* [$_{OBJ}$ *gli americani*] [$_{INF\text{-}A}$ *a restare a casa*].
"The fear of terrorism is recommending Americans to stay at home"

b. *La prima consiglia* [$_{OBJ}$ *la Freato*] [$_{INF\text{-}DI}$ *di telefonare*].
"The former recommends Ms. Freato to phone"

c. *Un giorno un medico consigliò* [$_{COMP\text{-}A}$ *a Dwight Eisenhower*] *[...]*[$_{INF\text{-}DI}$ *di fare ciclismo*].
"On day a physician recommended Dwight Eisenhower to bike"

The *Convincere* verbs are characterized by the high salience of the frames SUBJ#OBJ#FIN-*CHE* (5b) (including a direct object slot and a finite sentential complement introduced by *che*) and SUBJ#OBJ#COMP-*DI* (5c), that alternate with the SUBJ#OBJ#INF-*A* frame (5a). These verbs instead cannot occur neither with the frame SUBJ#OBJ#INF-*DI* nor with SUBJ#COMP-*A*#INF-*DI*, differently from the *Consigliare* verbs (5d-e):

(5) a. *Il terrorista convinse* [$_{OBJ}$ *la fidanzata*] [$_{INF\text{-}A}$ *a salire sull' aereo di El Al*].
"The terrorist convinced his fiancé to board on the El Al flight"

b. *Il grande sforzo adesso è convincere* [$_{OBJ}$ *gli italiani*] [$_{FIN\text{-}CHE}$ *che la partita di domani a Bari è importante*].
"The big effort now is to convince Italians that the match in Bari tomorrow is important"

c. *Il neo-capogabinetto ha convinto* [$_{OBJ}$ *Reagan*] [$_{COMP\text{-}DI}$ *dell'impossibilità di confermare l' incarico a Gates*].
"The new chief of staff convinced Reagan of the impossibility to confirm Gates' appointment"

d. **Gianni ha convinto* [$_{OBJ}$ *Maria*] [$_{INF\text{-}DI}$ *di visitare questo museo*].
"John convinced Mary to visit this museum"

e. **Gianni ha convinto* [$_{COMP\text{-}A}$ *a Maria*] [$_{INF\text{-}DI}$ *di visitare questo museo*].
"John convinced (*to) Mary to visit this museum"

In the *Consigliare* and *Convincere* classes the frame SUBJ#OBJ#INF-*A* is more marginal and marked than the other frames. Conversely, for the *Costringere* verbs, the class that encompasses the largest subset of the verbs selecting for SUBJ#OBJ#INF-*A*, this represents the most salient SCF. The verbs belonging to this large class are semantically similar to those in the *Indurre* and *Spingere* classes, for which the SUBJ#OBJ#INF-*A* SCF is also highly prototypical. The similarities are so close that we might even lump these three classes together. However, there is further distributional evidence supporting the decision of keeping them apart. For instance, the *Spingere* class is also

characterized by the frame SUBJ#OBJ#COMP-*CONTRO,* suggesting that these verbs express the idea of prompting somebody to act *against* somebody else. Some verb assignments are however not absolutely clear, consistently with the fact that the precise boundaries among these classes are hard to pin down. This is the case of a highly polysemous verb like *spingere* "push", which should presumably be assigned to more than one class. Indeed, the draft classification proposed in the Appendix as a flat list is an oversimplification. More realistically, semantic classes should have a multi-layered organization, with verbs potentially belonging to more than one class (cf. Levin 1993).

The classes that have been carved from the distributional verb profiles also greatly differ with respect to the semantic properties they share. Some classes, e.g. *Convincere* or *Autorizzare*, seem to be internally very coherent (with several near-synonym verbs), while others, for instance the *Costringere* class, present a higher degree of internal variation. A precise semantic characterization of these classes exceeds the limits of the present paper, but we can formulate some hypotheses about their meaning similarities. For instance, the semantic core of the *Costringere* class can be captured by assuming that their prominent SCF SUBJ#OBJ#INF-*A* is associated with a very complex "meaning component", which can roughly be represented with the following "event structure template" (cf. Rappaport Hovav & Levin 1998): *a Person* $X_{<SUBJ>}$ $ACTS_{<p>}$ *on a Person* $Y_{<OBJ>}$ *with the purpose that Y perfoms an act* $Z_{<INF\text{-}A>}$. This is indeed a general event schema interacting with other meaning components. For instance, verbs differ for the type of action that the subject performs on the object (formalized with the parameter <*p*> in the event structure template). This can be a coercive action (*costringere* "oblige") or a supportive one (*aiutare* "help"), within a much wider spectrum of possibilities. We can therefore model such a contrast by dividing the verb meaning into an *event structure component* – shared by a whole class of verbs - and a verb-specific *root component*, as proposed by Rappaport Hovav & Levin (1998). Another source of variation among verbs is represented by the *statistical salience* of the specific meaning component associated with the SCF SUBJ#OBJ#INF-*A*, possibly correlated with the relative prototypicality of this frame. In some verbs, this is the central (or dominant) aspect of meaning (e.g., *costringere* "oblige", *invitare* "invite", *spingere* "push", etc.). In some other verbs, this meaning component "competes" and/or interacts with other meaning components (cf. *consigliare* "recommend", *convincere* "convince", *mandare* "send", etc.). For instance, *convincere* has both a propositional interpretation, i.e. you act on somebody to make him/her believe that a certain state of affairs holds true, and an action interpretation, i.e. you act on somebody to make him/her perform some action. These interpretations are associated respectively with the SUBJ#OBJ#FIN-*CHE* and the SUBJ#OBJ#INF-*A* frames. The understanding of the complex relationships between different meaning components, together with

their distributional correlates, can indeed contribute to a better understanding of verb polysemy and its consequences for semantic verb classification.

4. Conclusions

In this paper, we have argued for the distributional approach as the correct method to pursue the goal of designing an empirically well-grounded semantic verb classification. We have also shown that the state of the art in computational linguistics can be used to turn this approach into an operative framework to build distributional profiles of verbs, representing the linguistic material from which verb classes can be carved. The small-scale experiment on Italian verbs goes exactly towards this direction. Now, we would like to conclude by raising some questions about the goal itself from which we have started, i.e. semantic verb classes. We will do this by reporting a quote from Levin (1993:17-18) that is too often overlooked in the literature on verb classification:

> The verb classes that are identified in this book should be "handled with care", since there is a sense in which the notion of "verb class" is an artificial construct. [...] The important theoretical construct is the notion of meaning component, not the notion of verb class.

Much work in computational linguistics and in lexical semantics has actually focused on the search for the best way to build classification schemes for verb meanings. However, there is a concrete risk that these efforts are actually missing the right goal. We have no doubt that verbs can be grouped into classes, since almost everything can be classified. The crucial issue are the features that we use to characterize the similarities among verbs supporting the classification. This is indeed the real missing aspect in the current research on verb classification: in fact, there is still little understanding of the meaning components, i.e. the semantic features, relevant to analyze verb meaning.[8] The distributional methodology - applied in this paper to Italian verbs - should therefore be used to address this specific goal, which is a necessary precondition for verb classification. Crucial improvements in this research can in fact be achieved only by reaching a better understanding of the complex interaction between the distributional patterns of verbs and the dimensions that govern their semantic space.

References

Attardi, Giuseppe & Felice Dell'Orletta. 2009. "Reverse Revision and Linear Tree Combination for Dependency Parsing". *Proceedings of NAACL-HLT 2009*, Boulder, Col.261-264.

Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston & Marco Mazzoleni. 2004. "Introducing the "la Repubblica" Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian". *Proceedings of LREC 2004*. Lisboa. 1771-1774.

Čulo, Oliver, Katrin Erk, Sebastian Padó & Sabine Schulte im Walde. 2008. "Comparing and Combining Semantic Verb Classifications". *Language Resources and Evaluation* 42:3.265-291.

Erk, Katrin. 2007. "A Simple, Similarity-Based Model for Selectional Preferences". *Proceedings of ACL*, Prague.216-223.

Evert, Stefan. 2008. "Corpora and Collocations". *Corpus Linguistics. An International Handbook* ed. by Anke Lüdeling & Merja Kytö, 1212-1248. Berlin: Mouton de Gruyter.

Fellbaum, Christiane, ed. 1998. *WordNet – An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Fillmore, Charles, Christopher Johnson & Miriam Petruck. 2003. "Background to Framenet". *International Journal of Lexicography* 16:3.235-250.

Hanks, Patrick. 1996. "Contextual Dependency and Lexical Sets". *International Journal of Corpus Linguistics* 1:1.75-98.

Hanks, Patrick & James Pustejovsky. 2005. "A pattern dictionary for natural language processing". *Revue Française de linguistique appliquée*.63-82.

Harris, Zellig S. 1954. *"Distributional Structure"*. *Word*, 10:2-3.146-62 [reprinted in Harris, Zellig S., 1970. *Papers in Structural and Transformational Linguistics*, Dordrecht: Reidel.775-794].

Kipper-Schuler, Karin. 2005. *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD dissertation, University of Pennsylvania.

Kipper-Schuler, Karin, Anna Korhonen, Neville Ryant & Martha Palmer. 2008. "A Large-Scale Classification of English Verbs". *Journal of Language Resources and Evaluation* 42:1.21-40.

Korhonen, Anna. 2009. "Automatic Lexical Classification - Balancing between Machine Learning and Linguistics". *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, Hong Kong.

Joanis, Eric, Suzanne Stevenson & David James. 2008. "A General Feature Space for Automatic Verb Classification". *Natural Language Engineering* 14:3.337-367.

Lapata, Mirella & Chris Brew. 2004. "Verb Class Disambiguation Using Informative Priors" *Computational Linguistic*s 3:1.45-73.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowsky, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas & Antonio Zampolli. 2000. "SIMPLE: A General Framework for the Development of Multilingual Lexicons". *International Journal of Lexicography* 13:4.249-263.

Lenci, Alessandro. 2008. "Distributional Semantics in Linguistic and Cognitive Research". *Italian Journal of Linguistics* 20:1.1-31.

Levin, Beth. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago, Ill.: University of Chicago Press.

Li, Janguo & Chris Brew. 2008. "Which Are the Best Features for Automatic Verb Classification". *Proceedings of ACL*, Columbus, Oh.434-442.

Light, Mark & Warren Greiff. 2002. "Statistical Models for the Induction and Use of Selectional Preferences". *Cognitive Science*: 26.269–281.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Language Processing*. Cambridge Mass.: MIT Press.

McCarthy, Diana. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD dissertation, University of Sussex.

Merlo, Paola & Eva Esteve Ferrer. 2006. "The Notion of Argument in Prepositional Phrase Attachment". *Computational Linguistics* 32:3.341-377.

Merlo, Paola & Stevenson Suzanne. 2001. "Automatic Verb Classification Based on Statistical Distributions of Argument Structure". *Computational Linguistics* 27:3.373-408.

Miller, George A. & Walter G. Charles. 1991. "Contextual Correlates of Semantic Similarity". *Language and Cognitive Processes* 6.1-28.

Pianta, Emanuele, Luisa Bentivogli & Christian Girardi. 2002. "MultiWordNet: Developing an Aligned Multilingual Database". *Proceedings of the 1st Global WordNet Conference*. Mysore.

Pustejovsky, James. 1995. *The Generative Lexicon*, Cambridge, Mass.: MIT Press.

Rappaport Hovav, Malka & Beth Levin. 1998. "Building Verb Meanings". *The Projection of Arguments* ed. by Miriam Butt & Wilhem Geuder, 97-134. Stanford, Cal.: CSLI Publications.

Roventini, Adriana, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini & Francesca Bertagna. 2000. "ItalWordNet: A Large Semantic Database for Italian". *Proceedings of LREC 2000*, Athens.783-790.

Schulte im Walde, Sabine. 2006. "Experiments on the Automatic Induction of German Semantic Verb Classes". *Computational Linguistics* 32:2.159-194.

Schulte im Walde, Sabine. 2009. "The Induction of Verb Frames and Verb Classes from Corpora". *Corpus Linguistics. An International Handbook* ed. by Anke Lüdeling & Merja Kytö, 952-972. Berlin: Mouton de Gruyter.

Sun, Lin & Anna Korhonen. 2009. "Improving Verb Clustering with Automatically Acquired Selectional Preferences". *Proceedings of EMNLP*, Singapore.638-647.

Vigliocco, Gabriella & David Vinson. 2007. "Semantic Representation". *The Oxford Handbook of Psycholinguistics* ed. by Gareth Gaskell, 195-215. Oxford: Oxford University Press.

Appendix - Verb classes

Costringere **verbs**

costringere "force", *invitare* "invite", *aiutare* "help", *obbligare* "oblige", *condannare* "condemn", *chiamare* "ask", *abituare* "get used", *sfidare* "challenge", *educare* "educate", *forzare* "force", *vincolare* "bind", *addestrare* "train", *richiamare* "recall", *designare* "designate", *pungolare* "goad", *rieducare* "re-educate", *allettare* "tempt", *istruire* "train", *incalzare* "ply", *predestinare* "predestinate", *sferzare* "incite", *sguinzagliare* "unleash", *deputare* "delegate"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - these verbs also typically occur with the SCF SUBJ#OBJ#COMP-*A*, with the OBJ slot selecting for the semantic classes PERSON or GROUP, and the COMP-*A* slot selecting for the semantic class ACT.

Indurre **verbs**

indurre "induce", *esortare* "exhort", *invogliare* "entice", *stimolare* "stimulate", *spronare* "goad", *orientare* "direct", *motivare* "motivate"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP. - these verbs also typically occur with the SCF SUBJ#OBJ#COMP-*A*, with the OBJ slot selecting for the semantic classes PERSON or GROUP, and the COMP-*A* slot selecting for the semantic class ACT; - these verbs also typically occur with the SCF SUBJ#OBJ#COMP-*VERSO*.

Spingere **- verbs**

spingere "push", *istigare* "instigate", *sospingere* "incite", *aizzare* "incite", *sensibilizzare* "sensitize" *incitare* "encourage"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - these verbs also typically occur with the SCFs SUBJ#OBJ#COMP-*CONTRO,* SUBJ#OBJ#COMP-*VERSO*, and SUBJ#OBJ#COMP-*A*.

Consigliare **- verbs**

consigliare "recommend", *ammonire* "admonish", *implorare* "implore", *supplicare* "beg"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - these verbs also typically occur with the SCF SUBJ#OBJ#INF-*DI*, with the OBJ slot selecting for the semantic classes PERSON or GROUP; - these verbs also typically occur with the SCF SUBJ#COMP-*A*#INF-*DI*, with the COMP-*A* slot selecting for the semantic classes PERSON or GROUP.

***Convincere* – verbs**

convincere "convince", *persuadere* "persuade"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - these verbs also typically occur with the SCF SUBJ#OBJ#FIN-*CHE*, with the OBJ slot selecting for the semantic classes PERSON or GROUP; - these verbs also typically occur with the SCF SUBJ#COMP-*DI* (es. *Gianni persuase Maria della necessità di partire* "John persuaded Mary about the necessity to leave")

***Autorizzare* verbs**

autorizzare "authorize", *sollecitare* "urge", *incoraggiare* "encourage", *delegare* "delegate", *incentivare* "stimolate", *abilitare* "qualify", *legittimare* "legitimate"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - this SCF alternates also with the SCF SUBJ#OBJ, selecting for the semantic classes ACT or EVENT: a. *Gianni ha autorizzato Maria a partire.* "John authorized Mary to leave" b. *Gianni ha autorizzato la partenza di Maria* "John authorized Mary's departure"

Scoraggiare - verbs

scoraggiare "discourage", *diffidare* "caution", *dissuadere* "dissuade", *sconsigliare* "not advise", *disincentivare* discourage
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - this SCF alternates also with the SCF SUBJ#OBJ#INF-*DA*, with the OBJ slot selecting for the semantic classes PERSON or GROUP, and the nominal infinitive: a. *Gianni ha dissuaso Maria a partire.* "John dissuaded Mary to leave" b. *Gianni ha dissuaso Maria dal partire* "John dissuaded Mary from leaving"

Portare - verbs

portare "bring", *destinare* "destinate", *mandare* "send", *condurre* "lead", *spedire* "send", *inviare* "send"
- the OBJ slot of the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes PERSON or GROUP; - the prototypical frame of these verbs is SUBJ#OBJ#COMP-*A*, with the OBJ slot selecting for the

semantic class ARTIFACT.

***Trascorrere* - verbs**

trascorrere “pass”, *destinare* “destinate”, *impegnare* “commit”, *cominciare* “begin”, *passare* “pass”, *impiegare* “commit”, *dedicare* “devote”, *iniziare* “begin”
- the OBJ slot of the SCF the SCF SUBJ#OBJ#INF-*A* typically selects for the semantic classes TIME, QUANTITY, or ACT.

SCF	LMI
SUBJ#COMP-*A*	104576.9044
SUBJ#0	66490.9049
SUBJ#COMP-*DA*	19680.8153
SUBJ#COMP-*IN*	17210.9291
SUBJ#SI#COMP-*A*	11577.3732
SUBJ#INF-*A*	9698.6682
SUBJ#COMP-*CON*	6963.6801
SUBJ#COMP-*SU*	3369.7406
SUBJ#COMP-*A*#COMP-*DA*	3115.0470
SUBJ#COMP-*ATTRAVERSO*	627.2822
SUBJ#SI#INF-*A*	507.2676

Table 1 – Syntactic profile for the verb *arrivare* "arrive"

Frame slots	Lexical sets	Selectional preferences
SUBJECT	*presidente* "president", *segretario* "secretary", *governo* "government" *proprietario* "owner", *datore* "employer", *medico* "doctor", *banca* "bank", *giornalista*, "journalist", etc.	PERSON GROUP
OBJECT	*decisione* "decision", *notizia* "news", *intenzione* "intention", *nome* "name", *variazione* "variation", *esito* "result", *disponibilità* "availability", *esistenza* "existence", *risultato* "result", *informazione* "information", *emozione* "emotion", *numero* "number", *senso* "sense", *dimissione* "dismissal", etc.	KNOWLEDGE ACT FEELING ATTRIBUTE COMMUNICATION STATE EVENT PROCESS
COMP-*A*	*autorità* "authority", *stampa* "press", *pubblico* "public", *lettore* "reader", *ministero* "ministry", *presidente* "president", *fisco* "tax office", *datore* "employer", *cliente* "customer", *sindacato* "trade union", *mercato* "market", etc.	PERSON GROUP

Table 2 – distributional profile for the SCF SUBJ#OBJ#COMP-*A* for *comunicare* "communicate"

COMP-*A*.PERSON	COMP-*A*.LOCATION
chiedere "ask"	*mettere* "put"
dare "give"	*rimettere* "restore"
affidare "entrust"	*portare* "carry"
offrire "offer"	*vedere* "see"
consegnare "deliver"	*colare* "sink"
inviare "send"	*buttare* "trash"
dire "say"	*collocare* "place"
raccontare "tell"	*mandare* "send"
rivolgere "turn"	*trovare* "find"
concedere "concede"	*accompagnare* "place"

Table 3 –verbs with the highest LMI for the classes PERSON and LOCATION as semantic types of the COMP-*A* slot in the SUBJ#OBJ#COMP-*A* frame

consegnare "deliver"	LMI	*mandare* "send"	LMI	*inviare* "send"	LMI
PERSON	6151.0897	GROUP	3825.7046	PERSON	3671.0614
GROUP	757.5376	NATURAL_OBJECT	431.5340	GROUP	924.1328
		LOCATION	358.5839	LOCATION	7.607
		PERSON	311.2284		

Table 4 – semantic preferences of the COMP-*A* slot in the SUBJ#OBJ#COMP-*A* frame of *consegnare* "deliver", *mandare* "send", and *inviare* "send"

* I am very grateful to Gabriella Lapesa for her precious help in carrying out the *LexIt* project. I also thank the two anonymous reviewers for their helpful comments. The usual disclaimers apply.

[1] Hence, the need for some sort of unification. The *Unified Verb Index* (`http://verbs.colorado.edu/verb-index`) is the first attempt at linking the major semantic classifications for English verbs.

[2] This definition comes from the FrameNet website: `http://framenet.icsi.berkeley.edu/`

[3] The extraction of distributional profiles has been carried out in collaboration with Gabriella Lapesa.

[4] See Merlo & Esteve Ferrer (2006) for a contribution in this direction.

[5] The issue of identifying the proper granularity of the noun semantic classes that best describe verb selectional preferences is still open, and has always been at the center of the debate in computational linguistics. Surely, these 24 classes are too broad to represent more subtle differences in verb semantic constraints. The work to extend the algorithm presented in this paper to a larger number of semantic classes is currently ongoing.

[6] The verb distributional profiles extracted from *La Repubblica* are freely accessible at this web site: `http://sesia.humnet.unipi.it/lexit`

[7] Interestingly, *sconsigliare* also occurs in the same corpus with the SUBJ#OBJ#INF-*DI* frame. The following example is almost a paraphrase of (2a):

(i) *I socialisti sconsigliano* [$_{\text{OBJ}}$ *Andreotti*] [$_{\text{INF-}DI}$ *di proseguire nel tentativo*].
"The socialist do not advise Andreotti to go on with his attempt"

[8] This point is also shared by Čulo et al. (2008), but they do not focus on the relationship between meaning features and verb distributional properties.