Building an Italian FrameNet through Semi-automatic Corpus Analysis

Alessandro Lenci, Martina Johnson, Gabriella Lapesa

University of Pisa, Department of Linguistics
Via Santa Maria 36, 56100 Pisa
alessandro.lenci@ling.unipi.it, martinaljohnson@gmail.com, g.lapesa@gmail.com

Abstract

In this paper, we outline the methodology we adopted to develop a FrameNet for Italian. The main element of novelty with respect to the original FrameNet is represented by the fact that the creation and annotation of Lexical Units is strictly grounded in distributional information (statistical distribution of verbal subcategorization frames, lexical and semantic preferences of each frame) automatically acquired from a large, dependency-parsed corpus. We claim that this approach allows us to overcome some of the shortcomings of the classical lexicographic method used to create FrameNet, by complementing the accuracy of manual annotation with the robustness of data on the global distributional patterns of a verb. In the paper, we describe our method for extracting distributional data from the corpus and the way we used it for the encoding and annotation of LUs. The long-term goal of our project is to create an electronic lexicon for Italian similar to the original English FrameNet. For the moment, we have developed a database of syntactic valences that will be made freely accessible via a web interface. This represents an autonomous resource besides the FrameNet lexicon, of which we have a beginning nucleus consisting of 791 annotated sentences.

1. Goals and Methodology

The long-term goal of our project is to create an electronic lexicon for Italian similar to the original English FrameNet (http://framenet.icsi.berkeley.edu). The aim of FrameNet is to characterize the meanings of words in terms of *semantic frames* (Fillmore, 1982; Fillmore, 1985) – schematic representations of the situations that characterize human experience, each constituted by a group of participants in the situation, or *Frame Elements* (FEs) – and to describe the possible syntactic realizations of the FEs for every word (Fillmore and Atkins, 1992; Fillmore et al., 2003). The focus of the lexicon is therefore on the syntactic-semantic interface. Italian FrameNet shares the aims of FrameNet, but in addition, it consistently integrates the statistical analysis of data from language corpora in its methodology, in an attempt to ground itself even more in distributional facts and the reality of linguistic usage.

The information necessary for the individuation of semantic frames is gathered by annotating corpus sentences with FEs (semantic roles) and syntactic information. In the Berkeley FrameNet, only a relatively small sample of sentences is annotated for every target word: this sample is selected manually, with the aim of making it representative of the word's most important syntactic-semantic combinatory possibilities. In Italian FrameNet, the same approach is being used, but the general distributional behavior of a verb is also taken into account and represented within the standard FrameNet format for Lexical Units (LU). In fact, we argue that it would be valuable to integrate the FrameNet development method with a rigorous and clearly defined methodology for the study of a word's syntactic distribution (in the spirit of Patrick Hanks' Corpus Pattern Analysis; Hanks and Pustejovsky (2005)), since this part of the annotation process is often criticized for relying too heavily on the individual annotator's intuition.

Various projects have focused on the creation of FrameNet for languages other than English, such as Spanish (Subirats, 2009), Japanese (Ohara, 2008) and German (Burchardt et al., 2009). Like the Berkeley FrameNet, these projects rely mainly on manual annotation. Besides this method, great interest exists nowadays in (semi)automatic approaches for bootstrapping FrameNets for new languages, typically employing methods derived from machine translation, or multilingual language processing in general (Chen and Fung, 2004; Tonelli et al., 2009). In this paper, we present the results of our distributional methodology to develop a FrameNet lexical resource for Italian, which draws on both kinds of approach. LU encoding (currently focusing on verbs) is performed manually, but the annotators' work relies on a wealth of syntactic and semantic information about verbal distributional behavior automatically extracted from a large corpus of Italian. Specifically, this includes the statistical distribution of a verb's subcategorization frames, as well as of the lexical and semantic preferences of each frame. Both these kinds of information are used by annotators to guide the construction of each LU.

2. Extracting Information about Verb Valence

In this section, we will describe the process of automatic extraction of distributional information on verbs from the *La Repubblica* Corpus (Baroni et al., 2004), a corpus of ca. 390 million word tokens of newspaper texts, which represents the source for the materials to be annotated in FrameNet.

The corpus was first lemmatized and part-of-speech tagged with the ILC-UniPi Tagger, and then dependency-parsed with DeSR, a state-of-the-art (88.6% Labelled Attachment Score) stochastic dependency parser (Bosco et al., 2009). The frequency distribution of verbs with various syntactic frames was then extracted from the parsed corpus. We applied Simple Log-Likelihood (Evert, 2008) to evaluate the correlation between:

1. verbs and syntactic frames;
2. slots and slot fillers, grouped by part of speech.

We calculated Simple Log-Likelihood (henceforth, LL) by applying the following formula *(O* represents the observed frequency of the pair, *E* its expected frequency):

$$Simple - ll = 2 \left(O \cdot log \frac{O}{E} - (O - E) \right)$$

The information on slot fillers is useful for identifying the specific lexical preferences of a verb occurring with a particular syntactic frame, i.e. to understand the prototypical nouns realizing a given FE in the semantic frame evoked by the verb.

2.1. Verb Selectional Preferences

The data described above were used to gain more insight about each verb's semantic preferences. Specifically, we wanted to assess the association strength between a predicate and the semantic types of its complements, as a further piece of information to be encoded in the LUs. To achieve this goal, we implemented the following variation of the algorithm described in Schulte Im Walde (2006), in order to assign selectional preferences to syntactic frames:

1. the co-occurrence frequency of each noun as a frame filler of a verb was uniformly divided among the different senses assigned to the noun in the Italian section of MultiWordNet (Pianta et al., 2002);

2. the sense frequency was then propagated up to the WordNet hierarchy to 25 mutually exclusive top-nodes (*Animal, Artifact, Act, Attribute, Food, Communication, Knowledge, Body, Event, Natural Phenomenon, Shape, Group, Location, Motivation, Natural Object, Person, Plant, Possession, Process, Quantity, Feeling, Substance, State, Time*). Thus, we obtained the joint frequency between each verb/frame/slot combination and the WordNet top-classes.

3. as an element of novelty with respect to Schulte Im Walde (2006), we calculated the LL association score between each verb/frame/slot combination and the 25 top-classes. This score was then used to represent the distribution of selectional preferences of a verb's frame-slot among the various semantic classes.

The set of semantic classes associated with the arguments of a verb constitutes its "semantic profile" (Alishahi and Stevenson, 2007), i.e. a representation of the semantic properties of its arguments. Such a profile has both a descriptive and predictive function because it represents the behaviour of the verb at the syntax-semantic interface but also provides a generalization that allows to make predictions about previously unseen arguments.

The data automatically extracted from the corpus – including verb preferences about subcategorization frames, lexical fillers and semantic classes – were then used to populate a mySQL database. A sample of the data is reported in Table 1 (in the actual database, frames, fillers and semantic classes come together with scores marking their statistical salience for the verb).

Slot	Fillers	Semantic Profile
Subject	presidente (*president*), proprietario (*owner*), segretario (*secretary*), medico (*doctor*), governo (*government*), ministro (*minister*), banca (*bank*), autorità (*authority*), amministratore (*administrator*), giornalista (*reporter*), azienda (*company*)	Person Group
Object	decisione (*decision*), intenzione (*intention*), notizia (*a piece of news*), nome (*name*), variazione (*variation*), dato (*datum*), esito (*result*), esistenza (*existence*), informazione (*information*), emozione (*emotion*), licenziamento (*dismissal*), risultato (*result*), sensazione (*feeling*), adesione (*adhesion*), data (*date*), elenco (*list*)	Knowledge Feeling Communication Act State Attribute Process Event
Comp-a	autorità (*authority*), stampa (*press*), pubblico (*public*), lettore (*reader*), ministero (*ministry*), datore (*employer*), fisco (*tax office*), cliente (*customer*), spettatore (*spectator*), sindacato (*trade union*)	Person Group

Table 1: *Comunicare* 'to communicate', syntactic frame: direct object + complement introduced by *a* 'to'

2.2. A Distributional Take on Polysemy

Our further step was an attempt to describe how polysemies are distributed in our corpus. We tried to model logical polysemy, "the ability of some words to appear in selectional contexts that are contradictory in type specification" (Pustejovsky, 2005). Typical examples of words showing this property are the following:

(1) a. Mary doesn't believe the book. INFO
b. John sold his books to Mary. PHYSOBJ

(2) a. I have my lunch in the backpack. FOOD
b. Your lunch was longer today than it was yesterday. EVENT

In the Generative Lexicon framework, words like *book* and *lunch* (see also *university*, *appointment*, *newspaper*) are given the status of complex types, because the concepts they express require an integration between types. The relation between the types integrated in a complex one is bidirectional and orthogonal: this is the reason why complex types are implemented as *dot-objects*.

As Pustejovsky (2005) pointed out, some verbs require complex types as their arguments. For example, *to read* selects a $physical.object \bullet information$ NP as a direct object, while *to fall* selects a $physical.object$ type for the same syntactic position. If a book ($physical.object \bullet information$ type) falls, however, it falls as a $physical.object$. If Paul reads a rumor about Mary, the rumor ($information$ type) undergoes a type shift: in order to be read it must have a physical manifestation.

The integration of dot-objects in our selectional preferences model is crucial to obtain a proper representation of argument structure properties. In order to add this information to our lexicon, we implemented the following algorithm:

1. for each verb, we select the frames and their fillers with $LL > 0$;
2. for each slot, we select the two semantic classes A and B with the highest LL, computed as described in section 2.1. The frequency of a word *w* filling the slot is then assigned:
   (a) to class *A,* if MultiWordnet assigns *w* to class *A* only, or to *A* and other classes not strongly associated with the slot.
   (b) to class *B,* if MultiWordnet assigns *w* to class *B* only, or to *B* and other classes not strongly associated with the slot.
   (c) to a potential dot-object *A-B,* if MultiWordNet assigns *w* to both *A* and *B*, or to *A* and *B* and to other classes not strongly associated with the slot.
3. As a result of the previous step, we obtained the joint frequency between each verb/frame/slot and class *A*, class *B*, and the potential dot-object *A-B*. Thus, we calculated the LL association between each slot and *A*, *B*, *A-B*.

The table in Appendix A reports an example of the results for the direct object in the transitive frame of the verbs *leggere* 'to read', *sfogliare* 'to flick through', *pubblicare* 'to publish', *bruciare* 'to burn', *vendere* 'to sell', *comprare* 'to buy'. These data provide a basis for some considerations. First of all, the wider pool of semantic classes selected by the verb (shown in the *Classes* column) provides an interesting representation of the degree of selectivity of some verbs with respect to others (in our example, *leggere*, *sfogliare* and *pubblicare* are more "picky" than *comprare*, *bruciare*, and *vendere*). The ranking among the first two classes and their dot-object allows us to identify the verbs that strongly select for complex types (again, *leggere, sfogliare* and *pubblicare* vs. *comprare, bruciare* and *vendere*).

The information concerning selectional preferences and inherent polysemy of arguments provides an interesting criterion for clustering syntactic positions showing similar semantic requirements. This kind of information is very useful both for the computational linguist and the FrameNet lexicographer. The computational linguist benefits from it when trying to carve verb classes from a corpus or to state argument realization generalizations. The FrameNet lexicographer will profit from this information in at least two moments of his/her annotation process: the semantic similarity between LUs can suggest that they belong to the same frame, and the identification of the most similar syntactic positions can indicate that these positions instantiate the same FE with respect to the frame evoked by the verb.

3. From Syntactic Frames to Semantic Frames

We are using the automatically acquired information on verb distribution described in Section 2. as an aid for the development of the Italian FrameNet lexicon. At the moment, the lexicon consists of 791 annotated sentences, featuring 6 verbs from the lexical domain of visual perception (*avvistare* 'to sight', *intravedere* 'to glimpse or make out', *notare* 'to notice', *osservare* 'to observe or watch', *sbirciare* 'to peek', and *scorgere* 'to glimpse or spot') and 9 frames, 3 related to perception (PERCEPTION EXPERIENCE, PERCEPTION ACTIVE, BECOMING AWARE), 5 to mental activity (AWARENESS, CATEGORIZATION, COMING TO BELIEVE, EXPECTATION, OPINION) and one to communication (STATEMENT). Annotation and LU encoding is carried out manually, through the *Berkeley FrameNet Desktop*. The Italian FrameNet database contains all the information found in the original FrameNet: a map of frame-to-frame relations, frame and Frame Element descriptions, and detailed reports on the syntactic realization of FEs for each LU. As a plus, it also records frequency information on the syntactic subcategorization frames of each LU, the prototypical lexical fillers of each frame's core FEs in relation to an LU, and their semantic types, expressed as a statistical distribution over WordNet top-classes and ranked according to their LL scores with that LU.

3.1. Using Verb Valence Information

The starting point of our method for encoding LUs is provided by the distributional data on verb valence properties, automatically extracted from *La Repubblica* as we explained in Section 2.. For each LU, we study the most frequent syntactic frames with which it occurs in order to select the ones that represent its most typical FE combinations (and their syntactic realizations). This approach allows us to register the most typical syntactic patterns for each LU and the differences in patterning between them. Table 2 shows the 16 most frequent syntactic patterns for *scorgere* and *sbirciare*.

As verbs of visual perception, both *scorgere* and *sbirciare* occur very frequently with a direct object expressing the perceived Phenomenon. However, almost all the other patterns for *sbirciare* feature a locative PP expressing the Direction of perception, while *scorgere*, on the other hand, occurs mostly with *in*-complements (which may express either the Direction or the Ground of perception) and *a*-complements (usually expressing the Place, but also Time), plus some non-locative PPs such as the *per*-complement (usually expressing Duration). This reflects a semantic difference between the two verbs: *sbirciare* always profiles

scorgere	**2783**	***sbirciare***	**491**
direct object	872	direct object	119
impersonal + no arguments	258	no arguments	71
no arguments	230	in (*in*)-comp.	47
dir. obj. + in (*in*)-comp.	229	da (*from*)-comp.	30
impers. + dir. obj.	176	tra (*between*)-comp.	20
in-comp.	93	dir. obj. + in-comp.	12
dir. obj. + a (*at/to*)-comp.	84	su (*on*)-comp.	12
impers. + in-comp.	52	a (*at/to*)-comp.	12
dir. obj. + su (*on*)-comp.	25	attraverso (*through*)-comp.	9
a-comp.	24	dietro (*behind*)-comp.	9
impers. + dir.obj. + in-comp.	22	verso (*toward*)-comp.	7
dir.obj. + a-comp. + in-comp.	21	dentro (*inside*)-comp.	7
impers. + a-comp.	21	impers. + no arguments	7
impers. + da (*from*)-comp.	20	dir. obj. + da-comp.	6
dir. obj. + che (*that*)-clause	17	con (*with*)-comp.	5
dir. obj. + per (*for*)-comp.	16	sotto (*under*)-comp.	5
...	...	...	...

Table 2: Syntactic patterns for *scorgere* and *sbirciare*

the direction of perception, while *scorgere* does not. This is part of the reason why they are assigned to different frames: *sbirciare* to PERCEPTION ACTIVE, where Direction is a core FE, and *scorgere* to PERCEPTION EXPERIENCE, where Direction is peripheral.

However, information on frequency is often necessary but not sufficient to determine which syntactic patterns are truly relevant for the semantic description of a word. For example, both *sbirciare* and *scorgere* occur with a number of fairly rare patterns which are typical of verbs of perception in Italian, and therefore relevant for the FrameNet lexicon. One of them, "direct object + *mentre* 'while'-clause", is exemplified in sentence (3) below. This pattern occurs only six times in *La Repubblica* with *scorgere*:

(3) **Ha scorto** l'ex presidente delle Ferrovie mentre faceva jogging in pigiama.
She glimpsed the ex-president of the railway company while he was jogging in his pajamas.

We find such rare but significant patterns through introspection, a study of the literature on the LUs we are encoding, and an in-depth analysis of corpus attestations. A manual study of text corpora conducted by competent annotators thus remains an important step in the construction of Italian FrameNet, just as it is in the original English FrameNet.

3.2. Using Information on Fillers

The Italian FrameNet method also takes into account the semantic types of the noun fillers of syntactic arguments. The distribution of the fillers is extracted from the corpus and automatically mapped onto WordNet top-classes, as we illustrated in Section 2. above. The semantic types of syntactic argument fillers are a necessary complement to information on the syntactic patterns occurring with an LU, both for the identification of the frame(s) it evokes, and for the individuation of the FEs that compose the frame.

Syntactic context is not always sufficient in order to identify the frame evoked by a word. This is one of the basic principles underlying Patrick Hanks' *Corpus Pattern Analysis* (CPA) that we decided to integrate in the Italian FrameNet development process. Hanks has noted that the combination of different semantic types in the same syntactic pattern often gives rise to different word senses: for example, *shoot* in the sentence *shoot a person* could conceivably be ambiguous, depending on whether the subject of the sentence is an armed attacker or a film director (Hanks and Pustejovsky, 2005, 68). The sense of the verb depends on the semantic type of the NP appearing as its subject. The same can be said in relation to FrameNet, by substituting the concept of "word sense" with "frame evoked by an LU". Here is an example based on our analysis of verbs of visual perception.

There is a wealth of studies on the fact that perception verbs (both in Italian and English) assume different interpretations depending on the syntactic constituent expressing the object of perception.[1] In Italian, this constituent can be an NP (4), a declarative *che* 'that'-clause ((7) and (8) below), or a construction specific to perception verbs, such as NP followed by an infinitive (5) or a pseudorelative clause (6) (see also the one exemplified in sentence (3) above). These constructions have their correspondents in English *that*-clauses and NPs followed by a naked infinitive or an *-ing* form, respectively.

(4) Il guardiacaccia **ha avvistato** per ben due volte l'orso bruno proprio nella sua valle.
The gamekeeper has sighted the brown bear not once, but twice in his own valley.

(5) Ride di cuore quando **sbircia** un fotografo inciampare nei fili delle cineprese.
He laughs heartily when he sees a photographer

[1] See for example Kirsner and Thompson (1976), Declerck (1981), Barwise (1981), Higginbotham (1983), and Guasti (1993).

trip/tripping on the camera cables.

(6) Il magistrato **scorge** un signore dall'aria distinta che si allontana in tutta fretta.
The judge glimpses a distinguished-looking man walk/walking away as quickly as possible.

In most studies, the focus has usually been on the difference between perception verb-specific complements and *che-* or *that*-clauses. It has been noted that, when a *che/that*-clause occurs as the complement of a perception verb, the verb no longer expresses a perceptual experience, but an act of deduction or reasoning based on perceivables. For example, if we say *"I see John playing tennis"*, we are relating a direct perceptual experience: we are in fact seeing John in the act of playing tennis at the moment of our utterance. If we say *"I see that John is playing tennis"*, on the other hand, this does not necessarily mean that we can see him playing (although this interpretation is also possible). We might have simply noticed that his racket and tennis shoes are missing from the usual place where he keeps them, and made a deduction based on that perceptual data.

Another way of expressing this is to say that a perception verb with a *che/that*-clause as complement assumes an epistemic meaning, i.e., it implies that the perceiver is aware of what is described in the complement. Not only does he see things, but he also has a conscious mental representation of them. The proposed reason for this is that *che/that*-clauses express a proposition, or, in intuitive terms, an epistemic content, whereas other possible constructions denote objects or events, i.e. entities in the world. It is possible to simply perceive an entity in the world, but a propositional content must be "held" mentally in a conscious way. Here are two examples featuring the verb *intravedere* 'glimpse', from *La Repubblica*:

(7) Con la tomografia abbiamo potuto **intravedere** che c'è una sedimentazione tra i due cervelli.
Thanks to the CAT scan, we could glimpse that there is some sedimentation between the two brains.

(8) In questi inizi del '90, già **intravediamo** che quelle novità sconvolgenti non sono nulla rispetto agli eventi che stanno per prodursi.
Now, at the beginning of the 90s, we can already glimpse that those shocking changes are nothing compared to the events that are about to unfold.

In sentence (7), *con la tomografia* 'thanks to the CAT scan' explicitly expresses the perceptual data on which the deduction expressed by the *che*-clause is based. *Intravedere* therefore retains a strong element of perceptual meaning, although the "object" of perception is actually a conclusion that must be believed or thought of. Sentence (8) makes no reference at all to physical perception: *intravedere* seems to have an exclusively epistemic interpretation in this case.

Verbs of visual perception with a *che/that*-clause therefore tend to lose their perceptual meaning and to acquire an interpretation that is related to mental activity instead (albeit a kind of mental activity that is based on perceivable data): their meaning becomes closer to that of epistemic verbs such as *know* and *believe*. Translating this in frame semantic terms, when a visual perception verb is followed by a *che/that*-clause, it no longer evokes a perception-related frame (e.g. PERCEPTION EXPERIENCE but a frame such as AWARENESS (typically evoked by the verbs *be aware, conceive, understand* in English) or OPINION (*believe, know, think*).

If we look at the data, however, we find that these frames are not evoked by verbs of visual perception only in presence of *che*-clauses: the same interpretation also emerges when the object is instantiated by NP whose noun filler denotes a non-perceivable entity. For example, observe the difference between the meaning of *intravedere* in sentence (9) and in sentences (10)-(12).

(9) Attraverso uno squarcio delle nuvole **intravedo** il maestoso ghiacciaio del vulcano spento Antisana. (*ghiacciaio* 'glacier': Natural Object)
Through an opening among the clouds I can glimpse the majestic glacier of the extinct volcano Antisana.

(10) Surin **intravede** in Jeanne le stesse passioni, gli stessi desideri dai quali è torturato lui. (*passione* 'passion', *desiderio* 'desire': Feeling)
Surin sees in Jeanne the same passions, the same desires that he himself is tortured by.

(11) Gli amici non avevano torto a **intravedere**, dietro le apparenze spettacolari, la saggezza di uno stoico antico. (*saggezza* 'wisdom': Knowledge)
His friends weren't wrong when they saw the wisdom of an ancient Stoic behind his spectacular appearances.

(12) Non solo Andreotti **intravede** una utilità nell'uso della comprensione verso Tripoli. (*utilità* 'usefulness': Attribute)
Andreotti isn't the only one who sees some usefulness in being sympathetic towards Tripoli.

In sentence (9), the NP appearing in the direct object slot belongs to the type Natural Object, which generally refers to visible entities: in this case, *intravedere* clearly expresses an event of physical perception. In sentences (10)-(12), on the other hand, the direct object is realized by nouns of the type Feeling, Knowledge, and Attribute, respectively. These classes are very often associated with "abstract" objects that it is impossible to perceive physically. As a consequence, *intravedere* practically loses its perceptual meaning; instead, it indicates that its subject (a Cognizer) holds an opinion based on what he or she perceives (for example, in sentence (10), Surin believes that Jeanne shares his same passions and desires, based on what he perceives of her). It is open to debate whether this sense of *intravedere* should be classified as a figurative interpretation or as an independent sense of the verb. We do not propose to answer this question at this time, but we simply point out that in these sentences *intravedere* cannot be said to evoke (only) the PERCEPTION EXPERIENCE frame, because it clearly evokes a frame related to mental activity (OPINION) as well.

This example shows how we use information on the semantic types of syntactic argument fillers, in concert with syntactic information, in order to identify the frame evoked by an LU during the encoding process. When extracting examples from the corpus, we do not just look at the syntactic pattern they instantiate (as described in Section 3.1.); we also look at the semantic types that appear in each syntactic slot. Then, we record the most significant combinations under the relevant frame. So, for example, the LU *intravedere* under PERCEPTION EXPERIENCE has among its possible syntactic-semantic combinations "direct object.NP = Natural Phenomenon", while *intravedere* under AWARENESS has "direct object.NP = Feeling/Knowledge/Attribute" – along with "*che*-clause", of course.

Information on semantic types is also necessary for the individuation of the FEs that compose a frame. This is because, even inside the same frame, the presence of different semantic types in the same syntactic slot may give rise to different interpretations of the syntactic argument, and consequently, the argument may be assigned different semantic roles.

For example, a PP introduced by *con* 'with' in the PERCEPTION ACTIVE frame can instantiate the Manner FE, the Instrument, or the Body Part used to perceive, depending on the semantic type of the noun that follows *con*: Feeling (as in sentence (13)), Artifact (sentence (14)), or Body Part (sentence (15)).

(13) Gli americani **osservano** [con crescente inquietudine]. (*inquietudine* 'disquiet': Feeling)
The American people keeps on watching, with growing disquiet.

(14) Una goccia di sangue **viene osservata** [con un microscopio tradizionale]. (*microscopio* 'microscope': Artifact)
A drop of blood is being observed with a traditional microscope.

(15) Il pilota si avvicinò al centro cittadino, **osservandolo** [con occhi fermi]. (*occhi* 'eyes': Body Part)
The pilot came closer to the city center, observing it with steady eyes.

When occurring in this particular syntactic construction, these FEs can be told apart exclusively by the semantic types of their fillers. In order to maintain this distinction, we select examples for annotation that do not just realize the relevant syntactic pattern (e.g. "direct object + *con*-comp."), but all relevant combinations of syntactic slots and filler types. We then record these combinations (which may be different for each LU) under the relevant FE: in this case, "direct object + *con*-comp. = Feeling" for Manner, "direct object + *con*-comp. = Artifact" for Instrument, and "direct object + *con*-comp. = Body Part" for Body Part.

4. Conclusions

In this paper, we presented the main features of the methodology we are currently adopting to develop a FrameNet for Italian. Its cornerstone is represented by the fact that LU creation and annotation is strictly grounded in distributional information automatically acquired from a large, dependency- parsed corpus. We claim that this approach allows us to overcome some of the shortcomings of the classical lexicographic method adopted to create FrameNet, by complementing the accuracy of manual annotation with the robustness of data that characterize the global distributional patterns of a verb. Besides, distributional analysis also allows us to enrich the format of FrameNet itself, extending LUs with information that provides a better characterization of a verb's behavior, i.e. corpus-derived selectional preferences and prototypical lexical fillers of a given syntactic pattern evoking a particular semantic frame. Both these types of information integrate the standard formal representation available in FrameNet, and can be extremely useful to address – within the FrameNet framework – key aspects of verb semantics, such as coercion phenomena, verb polysemy, etc. The database of syntactic valences will be made freely accessible via a web interface, and will represent a further autonomous resource besides the Italian FrameNet lexicon.

5. Acknowledgements

We would like to thank Collin Baker, Chuck Fillmore, and all the FrameNet team at ICSI, Berkeley CA, for granting us the use of the FrameNet Desktop and for giving us constant advice and support to start up the Italian FrameNet project.

6. References

Afra Alishahi and Suzanne Stevenson. 2007. A Cognitive Model for the Representation and Acquisition of Verb Selectional Preferences. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 41–48, Prague, Czech Republic.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. In *Proceedings of LREC 2004*, pages 1771–1774, Lisboa, Portugal.

Jon Barwise. 1981. Scenes and Other Situations. *The Journal of Philosophy*, 78(7):369–397.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice Dell'Orletta, and Alessandro Lenci. 2009. Evalita '09 Parsing Task: comparing dependency parsers and treebanks. In *Proceedings of EVALITA 2009*, Reggio Emilia, Italy.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009. Using FrameNet for the semantic analysis of the German: Annotation, representation, and automation. In H. C. Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 209–244. Mouton de Gruyter, New York.

Benfeng Chen and Pascale Fung. 2004. Automatic construction of an English-Chinese bilingual FrameNet. In *Proceedings of Human Language Technology conference/NAACL*, pages 29–32, Boston, MA.

Renaat Declerck. 1981. On the role of progressive aspect in nonfinite perception verb complements. *Glossa*, 15:83–114.

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin.

Charles J. Fillmore and Beryl T. (Sue) Atkins. 1992. Towards a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer and E.F. Kittay, editors, *Frames, fields and contrasts*, pages 75–102. Lawrence Erlbaum Associates, Hillsdale, NJ.

Charles J. Fillmore, Miriam R. L. Petruck, Josef Ruppenhofer, and Abby Wright. 2003. FrameNet in action: The case of attaching. *International Journal of Lexicography*, 16(3):297–332.

Charles J. Fillmore. 1982. Frame Semantics. In *Linguistics in the Morning Calm: Selected Papers from SICOL 1981*, pages 111–137. Hanshin, Seoul.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.

Maria Teresa Guasti. 1993. *Causative and perception verbs: A comparative study*. Rosenberg & Sellier, Turin, Italy.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

James Higginbotham. 1983. The Logic of Perceptual Reports: An Extensional Alternative to Situation Semantics. *The Journal of Philosophy*, 80(2):100–127.

Robert S. Kirsner and Sandra A. Thompson. 1976. The role of pragmatic inference in semantics: a study of sensory verb complements in English. *Glossa*, 10:200–240.

Kyoko Hirose Ohara. 2008. Lexicon, Grammar and Multilinguality in the Japanese FrameNet. In *Proceedings of LREC 2008*, pages 3264–3268, Marrakech, Morocco.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

James Pustejovsky. 2005. A Survey of Dot Objects. Technical report. Brandeis University.

Sabine Schulte Im Walde. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Carlos Subirats. 2009. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In H. C. Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 135–162. Mouton de Gruyter, New York.

Sara Tonelli, Daniele Pighin, Claudio Giuliano, and Emanuele Pianta. 2009. Semi-automatic Development of FrameNet for Italian. In *Proceedings of the FrameNet Workshop and Masterclass*, Milano, Italy.

Appendix A

Verb	Fillers	Classes	Classes & Dot-Objects
leggere (*to read*)	libro (*book*), giornale (*newspaper*), testo (*text*), articolo (*article*), lettera (*letter*), dichiarazione (*declaration*), romanzo (*novel*), pagina (*page*)	Communication (10371) Artifact (436) Time (28) Substance (17) Motivation (12)	Artifact-Communication (38194) Communication (18613) Artifact (-127)
sfogliare (*to flick through*)	pagina (*page*), margherita (*daisy*), giornale (*newspaper*), libro (*book*), album (*album*), catalogo (*catalogue*), rivista (*magazine*), volume (*volume*), quotidiano (*newspaper*), fascicolo (*issue*)	Communication (361) Artifact (340) Plant (118) Substance (48) Group (1.8)	Artifact-Communication (3232) Communication (568) Artifact (272)
pubblicare (*to publish*)	libro (*book*), foto (*picture*), articolo (*article*), lettera (*letter*), romanzo (*novel*), notizia (*a piece of news*), stralcio (*extract*), intervista (*interview*), testo (*text*), volume (*volume*), saggio (*essay*), disco (*record*)	Communication (5196) Artifact (888) Shape (2)	Artifact-Communication (16625) Communication (12978) Artifact (1363)
bruciare (*to burn*)	tappa (*stage*), bandiera (*flag*), tempo (*time*), sconfitta (*defeat*), auto (*car*), casa (*house*), cadavere (*corpse*), attualità (*current affairs*), incenso (*incense*), miliardo (*billion*), ettaro (*hectare*), copertone (*tyre*), cassonetto (*garbage bin*), corpo (*body*), candidatura (*nomination*), caloria (*calorie*), combustibile (*fuel*)	Artifact (354) Substance (279) Quantity (142) Natural Object (127) Plant (45) Natural Phenomenon (23) Time (20) Food (6) Body Part (6) Feeling (6) Location (1)	Artifact (3171) Substance (1145) Artifact-Substance (632)
vendere (*to sell*)	copia (*copy*), prodotto (*product*), milione (*million*), biglietto (*ticket*), azione (*stock*), quota (*share*), titolo (*bond*), pelle (*leather*), immobile (*building*), merce (*goods*), gioiello (*jewel*), pacchetto (*stake*), bene (*good*), disco (*record*), auto (*car*), sigaretta (*cigarette*), libro (*book*)	Artifact (2462) Substance (1332) Quantity (698) Possession (431) Food (401) Plant (74) Natural Object (16) Communication (14)	Artifact (13940) Artifact-Substance (7939) Substance (5722)
comprare (*to buy*)	azione (*stock*), biglietto (*ticket*), titolo (*bond*), casa (*house*), giornale (*newspaper*), auto (*car*), prodotto (*product*), libro (*book*), diritto (*right*), sigaretta (*cigarette*), disco (*record*), pacchetto (*stake*), dollaro (*dollar*), pane (*bread*), quota (*share*), appartamento (*flat*), marchio (*brand*), macchina (*car*), azienda (*company*), automobile (*car*)	Artifact (2236) Substance (796) Food (397) Possession (221) Group (61) Plant (21) Communication (8) Quantity (6) Process (4) Animal (2)	Artifact (10252) Artifact-Substance (4129) Substance (3771)

Table 3: *Leggere* ‘to read’, *sfogliare* ‘to flick through’, *pubblicare* ‘to publish’, *bruciare* ‘to burn’, *vendere* ‘to sell’, *comprare* ‘to buy’; Transitive frame, direct object.