LexIt: A Computational Resource on Italian Argument Structure

Alessandro Lenci[1], **Gabriella Lapesa**[2], **Giulia Bonansinga**[1]

[1]University of Pisa, Department of Linguistics, Via Santa Maria 36, 56100 Pisa, Italy
[2]University of Osnabrück, Institute for Cognitive Science, Albrechtstraße 28 49074 Osnabrück, Germany
alessandro.lenci@ling.unipi.it, glapesa@uos.de, giuliauni@gmail.com

Abstract

The aim of this paper is to introduce *LexIt*, a computational framework for the automatic acquisition and exploration of distributional information about Italian verbs, nouns and adjectives, freely available through a web interface at the address http://sesia.humnet.unipi.it/lexit. *LexIt* is the first large-scale resource for Italian in which subcategorization and semantic selection properties are characterized fully on distributional ground: in the paper we describe both the process of data extraction and the evaluation of the subcategorization frames extracted with *LexIt*.

Keywords: Distributional Methodology, Subcategorization, Selectional Preferences, Evaluation of Lexical Resources

1. Introduction

This paper introduces *LexIt*, a computational resource for the study of Italian verbs, nouns and adjectives at the syntax-semantics interface. The project that led to the realization of *LexIt* (today publicly available through a web interface at *http://sesia.humnet.unipi.it/lexit*) belongs to the longstanding research strand that aims at creating or extending lexical resources through the automatic acquisition of lexical information from corpora. In particular, the automatic acquisition of argument structure information is a widely explored topic that has recently experienced significant developments: extraction of subcategorization frames (Korhonen, 2002; Schulte im Walde, 2008), assignment of selectional preferences to arguments (Resnik, 1993; Light and Greiff, 2002; Erk et al., 2010), automatic induction of verb classes (Merlo and Stevenson, 2001; Schulte im Walde, 2006; Kipper-Schuler et al., 2008). Corpus-based information has been used to build lexical resources like VALEX for English (Korhonen et al., 2006), or LexSchem for French (Messiant et al., 2008). Such resources have represented an important reference points for the *LexIt* project.
In *LexIt*, syntactic and semantic properties of Italian predicates are characterized fully in terms of their *distributional profiles*, consisting of various types of statistical information describing their combinatory behavior. The *LexIt* methodology has the following main advantages:

- we go beyond the traditional distinction between argument and adjunct, which is often questionable and hard to turn into robust and clear-cut criteria;
- the extraction of subcategorization frames is totally unsupervised (we do not start from any given list of valence patterns). We instead discover the most salient frames using co-occurrence statistics of syntactic dependencies in parsed corpora;
- we integrate direct and inverse preferences of target predicates in subcategorization frames (Erk et al., 2010);
- the *LexIt* approach is the same for predicates belonging to different parts of speech.

LexIt is an open and parametrizable framework, which allows the researcher to explore argument structure as a function of many factors: source corpus, predicate part-of-speech, statistical indexes used to identify the most salient features of predicate argument structure (e.g., the most typical frames and arguments), semantic classes to model selectional preferences, etc.
This paper is structured as follows: in the first section we describe the process we applied to extract distributional profiles; the second section is meant to be a brief sketch of the *LexIt* query interface; in the third section we tackle the evaluation of part of the information acquired with *LexIt* (verb subcategorization frames); we will conclude by describing ongoing work for the extension of the resource.

2. Building Distributional Profiles

In *LexIt* each target lemma is associated with a distributional profile, an array of statistical information extracted from the corpus by applying state-of-the-art NLP methodologies and without any manual revision. The statistical information contained in each distributional profile is further articulated into:

- a *syntactic profile*, specifying the syntactic slots (subject, complements, modifiers, etc.) and the subcategorization frames with which the target predicate co-occurs;
- a *semantic profile*, composed by:
  - the *lexical set* of the most prototypical fillers realizing the syntactic slots;
  - the *semantic classes* characterizing the selectional preferences of syntactic slots.

To identify the most salient features of argument structures, each distributional feature is weighted with corpus frequency and the Local Mutual Information (LMI) score (Evert, 2008). The latter is a variant of the well-known Point-

wise Mutual Information and an approximation to the log-likelihood ratio measure that has been shown to be a very effective weighting scheme for sparse frequency counts. In *LexIt*, LMI is used to measure the association between verb and subcategorization frames (cf. section (4.)), frame slots and their lexical fillers, frame slots and semantic classes. The current version of *LexIt* contains information gathered from two different corpora: the *La Repubblica* (Baroni et al., 2004) corpus (ca. 331 millions tokens of newspaper articles) and the Italian section of *Wikipedia* (ca. 152 millions of tokens). In the pre-processing stage, the source corpora were tokenized, lemmatized and part-of-speech tagged with TANL (Text Analytics and Natural Language), a suite of modules for Italian Natural Language Processing developed by the University of Pisa and ILC-CNR. Dependency parsing was then performed with DeSR, a state-of-the-art (88.67% Labelled Attachment Score) stochastic dependency parser (Attardi and Dell'Orletta, 2009; Bosco et al., 2009).

2.1. Syntactic profiles

As anticipated in the introductory section, syntactic profiles in *LexIt* specify syntactic slots and subcategorization frames associated to target predicates (nouns, verbs and adjectives). Before describing the algorithm implemented for the extraction of syntactic profiles, it is worth defining what a Subcategorization Frame (henceforth SCF) is in *LexIt*. Clarification concerning this point is not only relevant from the theoretical point of view (e.g., with respect to the empirical contribution resources like *LexIt* can provide to explore the argument vs. adjunct distinction), but it is also crucial in the design of the resource, since it determines and constrains the users' search possibilities.

In *LexIt*, a SCF represents a pattern of syntactic dependencies headed by the target lemma. SCFs are synthetic labels formed by an unordered sets of slots, representing argument positions. Frame labels are composed by concatenating slot names with the symbol #. For example, the syntactic frame *complement introduced by a "to"* + *complement introduced by da "from"* is labeled as `comp-a#comp-da`. A set of slots is common to verbs, nouns and adjectives, while other slots that are part-of-speech specific. The argument slots that are common to all predicates are:

- complements: **comp-***, with * ranging over prepositions (e.g., *comp-a*, for the complement introduced by *a*, "to");
- infinitives: **inf-***, with * ranging over prepositions (e.g., *inf-di*, for the infinitive introduced by *di*, "of")
- finite clauses: **fin-***, with * ranging over subordinating conjunctions (e.g., *fin-che*, for the finite clause introduced by *che*, "that")

Among the verb-specific argument slots *LexIt* SCFs include:

- subjects (**subj**) and direct objects (**obj**);
- the zero argument construction (labeled as **subj#0**), corresponding to cases in which the only overtly realized argument is the is the subject (e.g., *Gianni piange*, "John cries");
- the reflexive pronoun **si** (e.g., *Gianni si lava*, "John washes himself");
- the predicative complement (**cpred** label) (e.g., *Anna sembra stanca*, "Ann seems tired").

Sentences in (1) exemplify cases in which the verb *dare*, "to give" occurs in the ditransitive frame (label:`subj#obj#comp-a`; verb-SCF joint frequency: 107,388; LMI: 327,656).

(1) a. *Gianni ha dato il libro a Maria* "Gianni gave the book to Mary"
b. *Gianni ha dato a Maria il libro* "Gianni gave Mary the book"
c. *Gianni ha generosamente dato a Maria il libro* "Gianni gave Mary the book generously"
d. *(Lui) ha dato il libro a sua madre piangendo* "(He) gave the book to his mother crying"

Examples of the the verb *rompere*, "to break" in the impersonal no-argument frame (label: `subj#si#0`: verb-SCF joint frequency: 1,980; LMI: 3,293) are the following:

(2) a. *Il vetro si è rotto*, "The glass broke"
b. *Il vetro si rompe facilmente*, "Glass breakes easily"

Cases in (1) and (2) show how the process of assignment of SCFs to target verbs abstracts from linear order of arguments, pro-drop and presence of verbal or adverbial modifiers. However, even if modifiers are not represented in the SCF labels, this type of information is retained in *LexIt* in dedicated *modifier* slots: **modadv** for adverbial modifiers, **modver** for verbal modifiers.

For target nouns, in addition to prepositional complements, infinitives and finite clauses, *LexIt* SCFs specify the zero argument construction, labeled as **0**. Adjectival modifiers are treated as the verbal and adverbial ones: they are not explicitly encoded in the frame labels, but the information concerning their fillers is stored in a dedicated slot, **modadj**. Examples in (3) show instances of the nominal SCF *complement introduced by di ("of")* + *complement introduced by in ("in, on, at")* in association with the target noun *colpo*, "shot" (label: `comp-di#comp-in`; noun-SCF joint frequency: 828; LMI: 1,396)

(3) a. *Un colpo di pistola in testa*, "A shot (of gun) on the head"
b. *Un brusco colpo di pistola in testa*, "A sudden shot (of gun) on the head"

Adjective-specific argument slots are the following:

- **pred**, containing the verbs with which the adjective occurs as a predicate (i.e., *essere* "to be", *apparire* "to appear");
- **mod-post**, containing the modified noun occurring *after* the adjective (i.e *grande uomo*, "a great man");
- **mod-pre**, containing the modified noun occurring *before* the adjective (i.e *uomo grande*, "a big man").

Cases in (4) and (5) exemplify, respectively, co-occurrence of the adjective *attento*, "careful" with the *predicative* frame (label:`pred`; adjective-SCF joint frequency: 5,242; LMI: 8,518) and with the *pre-adjectival head noun + complement introduced by "a"* (label:`mod-pre#comp-a`; adjective-SCF joint frequency: 305; LMI: 1,473).

(4) *Stai attento!* "Be careful!"

(5) *Un ministro attento a difendere...*, "A minister careful at protecting..."

From the linguistic point of view, adjective-specific slots have a different status from the slots shared with the other parts of speech. This is due to the specific nature of adjectives, which on the one hand govern prepositional and infinitival complements, and on the other hand occur as modifiers or predicates of other lexical items. Therefore, characterizing the distributional properties of adjectives requires to identify not only the type of slot they select (if any), but also the nouns they modify or the predicates they co-occur with. Our approach to adjectival subcategorization brings together, in the same SCF, direct selectional preferences (contstraints of predicates on their arguments) and inverse selectional preferences (preferences of arguments for their predicates). The potential of inverse selectional preference in improving the semantic representation extracted with distributional methodologies has already been explored from both the corpus-based and the cognitive point of view (Erk et al., 2010).

2.1.1. The SCF extraction algorithm

In order to extract the SCFs, we implemented the following algorithm separately for verbs, nouns and adjectives:

1. we automatically extracted from the parsed corpus the dependencies headed by a lemma belonging to the target part of speech (e.g., for verbs `subj`, `obj`, `comp-a`, etc.), plus other types of relevant information (e.g., the presence of the reflexive pronoun *si*). Each dependency represents a potential slot of the target lemma;

2. we computed the frequency of all the possible slot combinations (e.g. `subj#obj`, `subj#comp-a`, `subj#obj#comp-a`, etc.) attested in the corpus, and we selected the n most frequent ones as the potential SCFs for a given part of speech;

3. we extracted from the parsed corpus the co-occurrence frequency of each lemma with the selected SCFs;

4. the statistical salience of each SCFs with the target predicate was estimated in terms of LMI (cf. section (4.)). LMI proved to be particularly useful for the identification of the most prototypical SCFs for each predicate. Moreover, the application of LMI allowed us to downgrade mistaken frames due to parsing errors (e.g. PP-attachment).

Table 1 and 2 report the syntactic profiles associated with the verb *promettere* "to promise" and to the noun *promessa* "promise" in *La Repubblica*, ordered by decreasing LMI values.

SCF	Frequency	LMI
subj#inf-di	4,110	13,092.31
subj#fin-che	1,947	2,614.95
subj#obj	9,117	2,544.31
subj#obj#comp-a	1,623	1,079.03
subj#comp-a#fin-che	227	591,81
subj#comp-a#inf-di	247	457,19

Table 1: *promettere* 'to promise": syntactic profile

SCF	Frequency	LMI
inf-di	978	2,656.71
comp-da	285	417.23
comp-di#iinf-di	63	226.11
fin-che	100	154.71
comp-da_parte_di	26	52.15
comp-a#comp-da	18	27.67

Table 2: *promessa* "promise": syntactic profile

The comparison between the syntactic profiles of *attento* "careful" (table 3) and *verde* "green" (table 4) suggests that the integration of direct and inverse selectional preferences can provide significant contribution in capturing the distributional behavior of adjectives belonging to different classes: in this example, colors (prototypical case of intersective adjective) opposed to manner adjectives (more complex from the point of view of argument structure because of their reference to events).

SCF	Frequency	LMI
pred	5,242	8,518
pred#comp-a	1,378	5,706
mod-pre#comp-a	904	3,663
mod-pre#inf-a	305	1,473
pred#inf-a	318	1,027
mod-pre#inf-in	15	55
mod-pre#comp-in	55	46
mod-pre#comp-su	18	41
pred#comp-in	56	32
pred#inf-in	8	27

Table 3: *attento*,"careful": syntactic profile

2.2. Semantic profiles

Semantic profiles in *LexIt* are further articulated into lexical sets and selectional preferences of predicates over semantic classes of slot fillers.

The notion of lexical set (Hanks and Pustejovsky, 2005) defines the set of the words that typically occur with a target predicate in a given syntactic position. In *LexIt*, the lexical set assigned to each slot is composed by its fillers with LMI > 0 (computed over slot-filler co-occurrences).

Lexical sets available in *LexIt* are comparable to Sketch Engine's *word sketches*: "one-page automatic, corpus based

SCF	Frequency	LMI
mod-pre	10,474	6,883
pred#comp-come	4	1
mod-pre#comp-come	5	1
mod-pre#comp-con	4	1
mod-pre#inf-per	4	1

Table 4: *verde* "green": syntactic profile

summaries of a word grammatical and collocational behavior" (Kilgarriff et al., 2004). Similarly to Sketch Engine, *LexIt* uses grammatical patterns to describe every relation the target word participates in: both word sketches and syntactic profiles are defined in terms of a list of collocates occurring in each syntactic position. Patterns in *LexIt* and SketchEngine do not fully correspond, though: for instance, in *LexIt* the coordination relation is not considered, because not relevant for the study of argument structure. However, the main difference between SketchEngine's word sketches and *Lexit*'s lexical sets is in the possibility for *LexIt*'s user to rely on SCFs as a parameter to subgroup argument fillers (and corresponding semantic classes): the user can get a list of prototypical fillers for an argument slot of a target verb (for example, the subjects of the verb *dare*, "to give"), but can also subdivide this list among the different constructions in which the verb occurs (for example, by comparing the fillers of the subject of *dare* in the transitive and ditransitive construction).
Lexical sets were then used to gain more insight into the selectional preferences of the target predicates over the semantic classes of the words filling their argument slots. We implemented the following variation of the algorithm described in Schulte im Walde (2006):

1. the co-occurrence frequency of each noun as a slot filler in the lexical set associated to an argument slot was divided among the different senses assigned to the noun in the Italian section of MultiWordNet (Bentivogli et al., 2002).

2. the sense frequency was then propagated up the hierarchy, to 24 mutually exclusive top-nodes: 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. As a result, we obtained the joint frequency between each argument slot and the semantic classes.

3. as an element of novelty with respect to Schulte im Walde (2006), we calculated the LMI association between each argument slot and the 24 semantic classes.

Table 5 reports the semantic profile (lexical set and selectional preferences) for the complement introduced by *a* "to" of the verb *promettere*, "to promise" (source corpus: *La Repubblica*; between parentheses, the LMI values associated to lexical fillers and semantic classes).

Lexical Set	Semantic Classes
elettore "voter" (345,60)	PERSON (1541,61)
italiano "italian" (103,99)	GROUP (46,49)
vigilia "eve" (86,12)	ANIMAL (5,62)
moglie "wife" (83,48)	
presidente "president" (75,83)	
popolo "nation, people" (71,61)	
fine "end" (69,25)	
concittadino "fellow citizen" (67,66)	
cittadino "citizen" (66,10)	
paese "nation" (63,59)	

Table 5: *promettere* "to promise" - Complement introduced by *a* "to": semantic profile

Distributional semantic profiles have both a descriptive and a predictive function. On the one hand, lexical sets provide a sort of "snapshot" of the words co-occurring with a predicate 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 arguments, thereby making predictions about previously unseen slot fillers.

3. The resource and its interface

Currently, target predicates in LexIt are distributed as shown in table 6:[1]

POS	La Repubblica	Wikipedia.it
verbs	3,873	2,831
nouns	12,766	11,056
adjectives	5,559	

Table 6: Distribution of target predicates in Lexit (minimum frequency = 100)

The web interface allows the user to choose the part-of-speech (verb, noun or adjective) of the target lemma and the source corpus (*La Repubblica* or *Wikipedia.it*), and to query the database through five navigation paths:

- *by lemma* - to explore the distributional profile of a target lemma;
- *by syntactic frame* - to explore the lemmas that occurs with a target SCF;
- *by argument slot* - to explore the lemmas that occurs with a target slot;
- *by lexical filler and argument slot* - to explore the lemmas that occur with the target filler in a certain slot;
- *by semantic class* - to explore the lemmas that select for the target semantic class in a certain slot;

[1]The extraction of adjective profiles from *Wikipedia.it* is ongoing.

The possibility of combining different search parameters makes *LexIt* highly functional to address many types of research issues in computational linguistics and lexicography. Here, we will only discuss a small example of the impact of the choice of the source corpus on subcategorization results (Roland and Jurafsky, 1998). The comparison between the results of the search by target lemma in different corpora can be used to estimate the use of figurative language in newspaper articles (*La Repubblica*) in comparison with encyclopedic articles (*Wikipedia.it*). Tables 7 and 8 compare the semantic profiles for the subject of the verb *volare* "to fly" in La Repubblica and *Wikipedia.it*. The syntactic frame is the monovalent one (i.e., `subj#0`), the most prototypical one for the target verb in both corpora (between parentheses, the LMI values).

Lexical Set	Semantic Classes
insulto "insult" (813,26)	ARTIFACT (393,79)
parola "word" (744,20)	ANIMAL (392,40)
pugno "fist' (476,24)	COMMUNICATION (204,15)
schiaffo "slap" (305,30)	SUBSTANCE (98,36)
aereo "plane" (274,80)	ACT (40,88)
accusa "accuse" (198,16)	NATURAL OBJECT (33,28)
bottiglia "bottle" (174,44)	FOOD (28,39)
asino "donkey" (166,35)	QUANTITY (28,17)
calcio "kick" (143,134)	EVENT (27,19)
elicottero "helicopter"(134,29)	BODY PART (16,64)
sasso "stone" (129,91)	POSSESSION (10,99)
titolo "title" (128,94)	PLANT (4,92)

Table 7: What flies in *La Repubblica*?

Lexical Set	Semantic Classes
prototipo "prototype"(501,58)	ANIMAL (169,43)
aereo "plane" (134,19)	KNOWLEDGE (98,70)
notte "night" (110,33)	QUANTITY (28,78)
primo "first" (91,00)	ARTIFACT (24,01)
uccello "bird" (66,23)	NATUR. PHEN. (0.43)
volta "time" (41,50)	SUBSTANCE (0.37)
falco "hawk" (38,52)	
colombo "pigeon" (36,90)	
pilota "pilot" (31,83)	
letto "bed" (24,97)	
allodola "lark" (23,82)	
uomo "man" (22,42)	

Table 8: What flies in *Wikipedia.it*?

4. Evaluation

We evaluated the SCFs extracted with *LexIt*, following the methodology described in (Preiss et al., 2007) for English. The gold standard is represented by the valence patterns extracted from three manually developed Italian lexical resources:

- *Wörterbuch der Italianischen Verben* (Blumenthal and Rovere, 1998) (B&R) - Italian-German bilingual dictionary describing the meaning and valence properties of 1,729 Italian verbs;
- *Il Sabatini Coletti. Dizionario della Lingua Italiana* (Sabatini and Coletti, 2007) (S&C) - Italian monolingual dictionary in which verbs are marked with codes describing major valence patterns;
- PAROLE (Ruimy et al., 1998) - computational lexicon encoding the SCFs of 3,000 Italian verbs.

These resources greatly differ for the type and numbers of SCFs they describe. In B&R and S&C, SCFs are associated with verb senses, while *LexIt* links SCFs only to verb lemmas, abstracting away from specific SCF-meaning relations.

We restricted our evaluation only to verb SCFs extracted from *La Repubblica*. We randomly selected 100 verbs from the 3,873 verbs for which *LexIt* extracts SCFs from that corpus. Since some of these verbs did not occur in one or more of the gold standards, we repeated the random sampling until all the 100 verbs were attested in each of the three dictionaries above (min. freq. 429 *miscelare* "mix"; max. freq. 830,903 *dire* "say").

Given the great differences in the way valence patterns are represented in each gold standard and in *Lexit*, checking which extracted frames also appear in the lexical resources is not a straightforward operation. Therefore, for each *LexIt* SCF, we manually verified whether it was attested in the gold standard. As a general strategy, we did not consider as error any mismatch between *LexIt* and the gold standard due to the inherent design features of the extraction process. In some cases, gold standard frames make distinctions that exceed the scope of *LexIt*. For instance, the three gold standard resources assume some kind of argument-adjunct distinction. That is, coded valence patterns report core verb arguments, but ignore possible adjuncts or circumstantial slots. Since this dichotomy is not captured in *LexIt* "by design", we regarded a *LexIt* frame like `subj#obj#comp-in` as a true positive even if the gold standard only reports a subject-object frame, provided that `comp-in` is a possible adjunct phrase for that verb. We only excluded those prepositional slots that were clearly wrong, typically because of PP-attachment mistakes by the parser. Another example is provided by the Italian reflexive pronoun *si*. The SCFs in PAROLE encode a very fine-grained distinction between different uses of *si* in Italian, such as true reflexive constructions, impersonal uses, pronominal intransitives (in fact, for some verbs *si* is just an intransitivity marker, like with *rompersi* "break (inchoative)"), etc. Capturing these differences goes well beyond the expressive capability of *LexIt*, and actually exceeds the the state of the art of parsing systems too. As a matter of fact, *LexIt* only distinguishes verb frames containing the reflexive pronoun (e.g., `subj#si#0`), from those that do not contain it (e.g., `subj#0`). Consistently, we decided not to count more fined-grained distinctions as false negatives in the present evaluation.

In other cases, *LexIt* SCFs make more subtle distinctions than those found in valence dictionaries. For instance, S&C does not distinguish the specific preposition heading frame slots (e.g., it only contains a generic frame `sogg-v-arg-prep.arg`, without information about the type of preposition). Instead, *LexIt* considers the

preposition heading a slot as a distinctive feature for frames (e.g., `subj#obj#comp-in` and `subj#obj#comp-su` are regarded as two different SCFs). In these cases, we regarded the *LexIt* SCF as correct, if the gold standard contains a frame with a prepositional slot, and the *LexIt* preposition is acceptable for the given frame and slot. To decide the acceptability of the prepositions we looked at the example sentences in the lexical resources (if available) or at corpus examples. Similarly, there are cases in which a dictionary only reports a subset of the possible range of preposition heading a slot (e.g., it can specify a locative preposition *a* "at" but not other equally locative prepositions such as *in* "in" or *su* "on"). If *LexIt* has a frame containing a preposition not attested in the gold standard, but with the same function as those therein specified, we judged the *LexIt* frame correct.

The standard practice to evaluate SCF extraction is to filter extracted frames with respect to some statistical score to exclude possibly "noisy" frames due to tagging and parsing errors (Korhonen, 2002). In particular, only SCFs with a score above a certain empirically determined threshold are evaluated. We follow the same procedure and we actually adopted two types of scores to rank the *LexIt* frames:

- Maximum Likelihood Estimation (MLE) - this is the same type of score used by (Preiss et al., 2007) and (Messiant et al., 2008) and corresponds to the relative frequency of a SCF i with a verb j calculated as follows:

$$rel\_freq(scf_i, v_j) = \frac{f(scf_i, v_j)}{f(v_j)}$$

  where $f(scf_i, v_j)$ is the joint frequency of the verb j with the SCF i, and $f(v_j)$ is the verb frequency in the corpus;

- Local Mutual Information (LMI) - as we said in section (2.), the prototypicality of verb frames, slot fillers and semantic classes is estimated in *LexIt* with the LMI association score. The LMI between a SCF i with a verb j is calculated as follows:

$$LMI(scf_i, v_j) = f(scf_i, v_j) * log_2 \frac{p(scf_i, v_j)}{p(scf_i)(p(v_j)}$$

  The LMI actually corresponds to the verb-SCF joint frequency weighted with Pointwise Mutual Information between the verb and the SCF.

For each verb we calculated type precision, type recall and F-measure over each of the three gold standards at increasing thresholds of MLE and LMI scores. That is, for increasing values of k, we considered only the SCFs whose MLE or LMI score was bigger than k. The figures (1) and (2) plot the F-measure (averaged among the 100 test verbs) with respect to different MLE and LMI thresholds, computed over the three gold standards resources.

The tables (9) and (10) report the best F-measure over the various gold standards. For MLE, the best scores have been obtained with a relative frequency threshold between 0.01 and 0.02. As can be seen in Figure (2), the F-measure dynamics with LMI is more complex, and best scores are typically obtained with a threshold between 100 and 200. However, LMI scores seems to score better than MLE, especially with respect to precision. This also confirms the utility of LMI to filter out noisy frames. Both measures instead produce a much higher recall than precision. This is also consistent with the unsupervised approach to SCF extraction adopted by *LexIt*, as well as by possible mistakes due to corpus pre-processing.

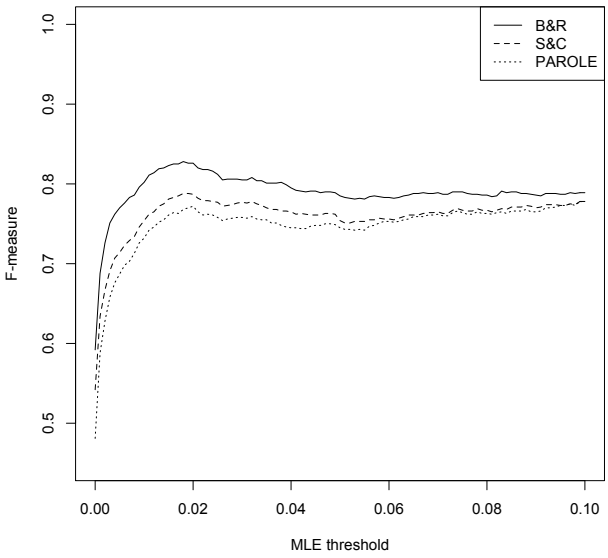

Figure 1: SCF F-measure and MLE threshold

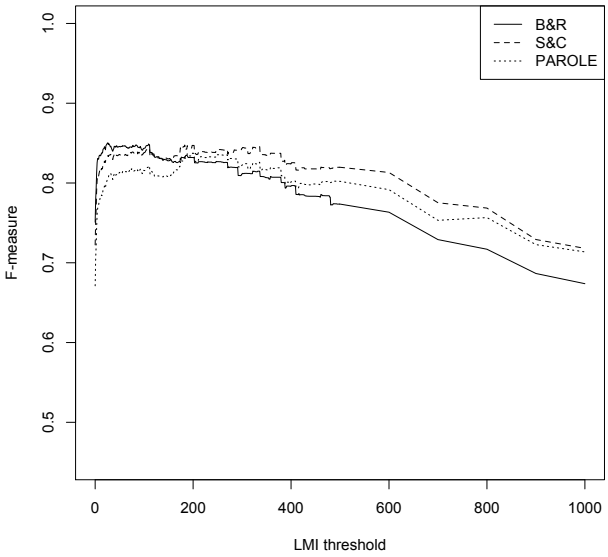

Figure 2: SCF F-measure and LMI threshold

Gold standard	Precision	Recall	F-measure
B&R	0.78	0.91	0.82
S&C	0.69	0.95	0.78
PAROLE	0.69	0.97	0.78

Table 9: Top scores with MLE thresholds

Gold standard	Precision	Recall	F-measure
B&R	0.82	0.92	0.85
S&C	0.80	0.95	0.85
PAROLE	0.77	0.96	0.84

Table 10: Top scores with LMI thresholds

5. Conclusions and Future Work

In this paper, we described the extraction of distributional profiles for Italian verbs, nouns and adjectives that have been used to populate *LexIt*, a corpus-derived lexical resource for Italian freely accessible via a Web interface. The first evaluation of *LexIt*, focussed on verbs SCF, has proved the high accuracy of the information extracted with *LexIt*: precision and recall values are indeed comparable with other state-of-the-art corpus-derived valence lexicons. Besides refining the methodology for SCF extraction, ongoing work on *LexIt* includes:

- extracting distributional information from domain corpora;
- integrating semantic profiles with information concerning argument polysemy and semantic roles;
- adding distributional profiles for multiword expressions;
- carrying out semi-automatic classifications of Italian verbs using the *LexIt* distributional profiles.

6. References

G. Attardi and F. Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL-HLT 2009*, pages 261–264, Boulder, USA.

M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. 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.

L. Bentivogli, E. Pianta, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the 1st International WordNet Conference*, pages 293–302, Mysore, India.

P. Blumenthal and G. Rovere. 1998. *Wörterbuch der italienischen Verben*. Ernest Klettverlag, Stuttgart.

C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, and A. Lenci. 2009. Evalita'09 parsing task: comparing dependency parsers and treebanks. In *Proceedings of EVALITA 2009*, Reggio Emilia, Italy.

K. Erk, S. Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

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

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 2:63–82.

A. Kilgarriff, P. Rychlý, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. In *Proceedings of EURALEX 2004*, pages 105–116, Lorient, France.

K. Kipper-Schuler, A. Korhonen, R. Neville, and M. Palmer. 2008. A large-scale classification of English verbs. *Journal of Language Resources and Evaluation*, 42(1):21–40.

A. Korhonen, Y. Krymolowski, and T. Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC 2006*, Genova, Italy.

A. Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge.

M. Light and W. Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 87:1–13.

P. Merlo and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

C. Messiant, A. Korhonen, and T. Poibeau. 2008. LexSchem: A large subcategorization lexicon for French verbs. In *Proceedings of LREC 2008*, Marrakech, Morocco.

J. Preiss, T. Briscoe, and A. Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL 2007*, pages 912–919, Prague, Czech Republic.

P. Resnik. 1993. *Selection and information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania.

D. Roland and D. Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of COLING-ACL*, pages 1122–1128, Montréal, Canada.

N. Ruimy, O. Corazzari, E. Gola, A. Spanu, N. Calzolari, and A. Zampolli. 1998. LE-PAROLE project: The Italian syntactic lexicon. In T. Fontanelle and Hiligsmann P., editors, *Proceedings of EURALEX 1998*, volume I, pages 259–269, Liège.

F. Sabatini and V. Coletti. 2007. *Il Sabatini-Coletti: dizionario della lingua italiana*. Rizzoli-Larousse, Milano.

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

S. Schulte im Walde. 2008. The induction of verb frames and verb classes from corpora. In Lüdeling A. and Kytö M., editors, *Corpus Linguistics: An International Handbook*, chapter 61. Mouton de Gruyter, Berlin.