LexFr: Adapting the *LexIt* Framework to Build a Corpus-Based French Subcategorization Lexicon

Giulia Rambelli*, Gianluca E. Lebani*, Laurent Prévot†, Alessandro Lenci*

* Computational Linguistics Laboratory, University of Pisa, via Santa Maria, 36, Pisa (Italy)
† Aix-Marseille University, avenue Pasteur, 5, Aix-en-Provence (France)
g.rambelli1@studenti.unipi.it, gianluca.lebani@for.unipi.it, laurent.prevot@lpl-aix.fr, alessandro.lenci@unipi.it

Abstract

This paper introduces *LexFr*, a corpus-based French lexical resource built by adapting the framework *LexIt*, originally developed to describe the combinatorial potential of Italian predicates. As in the original framework, the behavior of a group of target predicates is characterized by a series of syntactic (i.e., subcategorization frames) and semantic (i.e., selectional preferences) statistical information (a.k.a. *distributional profiles*) whose extraction process is mostly unsupervised. The first release of *LexFr* includes information for 2,493 verbs, 7,939 nouns and 2,628 adjectives. In these pages we describe the adaptation process and evaluated the final resource by comparing the information collected for 20 test verbs against the information available in a gold standard dictionary. In the best performing setting, we obtained 0.74 precision, 0.66 recall and 0.70 F-measure.

Keywords: Automatic Lexical Acquisition, Subcategorization, Selectional Preferences, Evaluation of Lexical Resources

1. Introduction

From the very beginning of the Natural Language Processing (NLP) enterprise, a crucial research topic has been the development of (semi-) automatic methods to build or enrich lexical resources. It didn't take long to realize that the accuracy advantage of the traditionally hand-built resources were counterbalanced by the higher cost-effectiveness and flexibility of the automatic methods, features that come in handy especially for languages, domains of topics for which hand-built resources are not available or are just too limited in their scope.

Among the several kinds of information that can be included in a lexicon, the description of linguistic entities at the syntax-semantic interface proved to be useful for many traditional Natural Language Processing tasks such as word-sense disambiguation, machine translation, knowledge extraction (Schulte im Walde, 2009; Korhonen, 2009), so that the automatic acquisition of argument structure information is a long-standing topic in computational linguistics. By embracing the theoretical assumption that the semantics of a predicate and the morpho-syntactic realization of its arguments are intimately related (Levin, 1993; Bresnan, 1996; Roland and Jurafsky, 2002; Levin and Rappaport Hovav, 2005), the last thirty years have witnessed the development of automatic methods for the identification of verb subcategorization frames (Korhonen et al., 2006; Messiant et al., 2008; Schulte im Walde, 2009), selectional preferences (Resnik, 1996; Light and Greiff, 2002; Erk et al., 2010) and diathesis alternation (McCarthy, 2001).

The literature reports few examples of automatically built, wide coverage, lexica encoding this information, a.k.a. combinatorial lexica, among which notable mentions include *VALEX* for English verbs (Korhonen et al., 2006), *LexSchem* from French verbs (Messiant et al., 2008) and *LexIt* for Italian verbs, nouns and adjectives (Lenci et al., 2012). These resources represent a reference point for the work presented in these pages, where we investigated the possibility to automatically extract distributional information about French predicates by adapting an existing framework, *LexIt*. This led to the realization of *LexFr*, an automatically built French lexicon describing the syntactic and semantic properties of the argument structures of 2,493 verbs, 7,939 nouns and 2,628 adjectives. As in the original framework, the behavior of a group of target predicates is characterized by a series of statistical information (a.k.a. *distributional profiles*) whose extraction process is mostly unsupervised.

These pages are organized as follows: we begin with a quick review of the existing French conbinatorial lexica; sections 3 and 4 describes the *LexIt* framework and its adaptation to French; section 5 describes the resulting resource, which we evaluated by comparing 20 test verbs against a gold standard dictionary; we will conclude by reporting some possible improvements and ongoing research.

2. French Combinatorial Lexica

With few exceptions, most of the currently available French combinatorial lexica are hand-built. These resources differ in terms of coverage, granularity, formalization, argument/adjunct identification and reference linguistic theory (Sagot and Danlos, 2012). Examples include:

- Maurice Gross' *Lexicon-Grammar* (Gross, 1975): an electronic dictionary containing a systematic description of syntactic and semantic properties of verbs, nouns and adjectives. This lexicon is organized into tables, each table describing the syntactic and semantic properties of a particular syntactic construction and listing the lexical items showing each construction. Only part of this resource is publicly available.
- the *Lexique des Verbes Français* (Dubois and Dubois-Charlier, 1997) is a dictionary encoding several kinds of semantic and syntactic information pertaining to 12,130 verbs: class, meaning, linguistic domain,

conjugation and auxiliary, syntax of verb (transitive/intransitive, subcategorization frame, etc.), morphological derivation, sample sentences. A crucial aspect of this resource is its being centered around the notion of syntactically-characterized semantic verb class, thus exploiting the idea that the syntactic behavior of a predicate reflects some key aspects of its semantics (Levin, 1993). Accordingly, the thesaurus-like structure of this resource is organized over 5 levels of classification, the more general one distinguishing between 14 semantic classes.

- *Dicovalence* (van den Eynde and Mertens, 2010) is a valency lexicon containing information for more than 3,700 verbs. It is based on the pronominal approach (van den Eynde and Blanche-Benveniste, 1978), a research method that treats pronouns as semantic primitives due to the purely linguistic nature and finite inventory of this lexical class. Accordingly, in this resource valence slots are characterized by the set of accepted pronouns, which subsume the possible lexicalizations of that slot.

- The *Lexique des Formes Fléchies du Français*, a.k.a *Lefff* (Sagot, 2010), is a semi-automatically built morphological and syntactic lexicon. The several manual validation, correction and extension steps needed to build this resource led some authors to describe it as "an automatically acquired morphological lexicon [...] which has been manually supplemented with partial syntactic information" (Messiant et al., 2008). Version 3.0.1 of this resource describes more than 110k lexical elements, among which 6,825 verbs, 37,530 nouns and 10,483 adjectives. Its lexical framework, *Alexina* (Architecture pour les LEXiques INformatiques et leur Acquisition), has been successfully exploited to create *Lefff*-like resource in other languages such as Italian and Dutch.

Well known issues with hand-built resources include their coverage, their laborious and time-intensive population and maintenance, as well as the lack of statistical information concerning the described phenomena (e.g., the likelihood of a certain subcategorization frame for a given verb). Projects like *TreeLex* (Kupść and Abeillé, 2008) and *LexSchem* (Messiant et al., 2008) have been developed to plug such a gap.

TreeLex is a valence lexicon for French verbs (Kupść and Abeillé, 2008) automatically built from a newspaper treebank composed by 5 years of the French newspaper 'Le Monde'. In this resource, the combinatorial behavior of approximately 2,000 verbs is described by means of an inventory of 160 subcategorization frames. Subsequent works by the same authors tested the possibility to expand *TreeLex* to report the cooccurrence statistics of approximately 2,000 adjectives into 40 subcategorization frames (Kupść, 2009). To the best of our knowledge, such an extension has not been implemented in the publicly available version of the resource. As pointed out by Messiant et al. (2008), a major drawback of the *TreeLex* framework lays in its reliance on manual effort, namely in the need for manual correction of the output of the automatic module.

LexSchem has been the first automatically built lexical resource characterizing the subcategorization behavior of a large set of French verbs (Messiant et al., 2008). This information has been extracted by using *ASSCI* (Messiant, 2008), a subcategorization frames acquisition system whose main task is to extract all the patterns for each target verb and exploit a MLE-based strategy (see section 5) to identify the more plausible set of subcategorization frames. By applying ASSCI to a newspaper corpus composed by 10 years of the French newspaper 'Le Monde', 336 subcategorization frames have been isolated and used to describe the combinatorial behavior of 3,297 French verbs. The goodness of the *LexSchem* framework has been tested by comparing the entries for 20 test verbs against a gold standard dictionary, thus showing 0.79 precision, 0.55 recall and 0.65 F-measure.

Apart from their cost-effectiveness, a crucial advantage of the information available in resources like *TreeLex* and *LexSchem* over the traditional, hand-built lexica lays in their encoding of frequency-based information such as the joint frequency of each verb and subcategorization frame. Major drawbacks of these resources, when compared with frameworks like *LexIt* (see below), are their being limited to verbal valencies and the lack of semantic information like selectional preferences and semantic roles.

3. The *LexIt* Framework

LexIt (Lenci et al., 2012) is both a computational system, as well as an on-line database[1] containing distributional information of Italian verbs, nouns and adjectives. In this resource, the linguistic annotation available in a parsed corpus is processed in order to describe the combinatorial behavior of a set of target predicates by means of *distributional profiles*, a data structure that blends different kinds of statistical information, and that is further articulated into *syntactic profiles* and *semantic profiles*.

3.1. Syntactic Profiles

The *syntactic profile* of a given target predicate characterizes the statistical association between the predicate and all the syntactic arguments (a.k.a. *syntactic slots*: e.g., subject, object, complements, modifiers, etc.) and subcategorization frames (SCFs) with which it occurs. In *LexIt*, arguments and adjuncts are treated alike, so that a SCF represents an unordered pattern of syntactic dependencies headed by the target predicate, and it is labeled by concatenating its atomic slots names with the symbol "#". For instance, the simple transitive SCF composed by a subject and an object is marked as `subj#obj`.

Due to the different syntactic behavior of parts of speech, the inventory of syntactic slots in *LexIt* is mostly PoS-dependent, except for the following argument slots that are common to all predicates:

- complements are labeled as $\texttt{comp}_*$, with "*" ranging over prepositions: e.g. $\texttt{comp}_{\grave{a}}$, for the complement introduced by the preposition *à* ("to");

[1] available at `http://lexit.fileli.unipi.it/`

- infinitives are labeled as $\texttt{inf}_*$, with "*" ranging over prepositions: e.g. $\texttt{inf}_{\texttt{de}}$ for the infinitive introduced by the preposition *de* ("of");
- finite clauses are labeled as $\texttt{fin}_*$, with "*" ranging over prepositions: e.g. e.g. $\texttt{inf}_{\texttt{que}}$ for the infinitive introduced by the preposition *que* ("that").

Table 1 lists the PoS-dependent argument slots that can appear in a SCF. Syntactic dependencies that are extracted but that are not represented in the SCFs are those involving verbal (modver), adverbial (modadv) and adjectival (modadj) modifiers.

PoS	Label	Argument Slot
verbs	0	zero argument construction
	subj	subject
	obj	direct object
	se	reflexive pronoun
	cpred	predicative complement
nouns	0	zero argument construction
adjectives	pred	the verb bearing the target as a predicate
	mod-post	modified noun occurring after the target
	mod-pre	modified noun occurring before the target

Table 1: PoS-specific SFC argument slots in *LexIt*

The following examples shows some possible argument realizations of the verb *promettre* ("to promise"):

1. *Je promis (à Anne) de faire l'impossible*
   $\texttt{subj\#comp}_{\texttt{à}}\texttt{\#inf}_{\texttt{de}}$
   "I promised to Anne to do the impossible"
2. *Le soldat promet la vie sauve à son camarade*
   $\texttt{subj\#obj\#comp}_{\texttt{à}}$
   "The soldier promised his comrade to save his life"
3. *Jean se promit de se corriger*
   $\texttt{subj\#se\#inf}_{\texttt{de}}$
   "John promised to correct himself"

3.2. Semantic Profiles

The *semantic profile* of a given target predicate enriches the information contained in its *syntactic profile* by characterizing:

- the *lexical set* of the typical lexical items filling each syntactic slot;
- the *semantic classes* of these same lexical items, to characterize the selectional preferences of syntactic slots.

Following Hanks and Pustejovsky (2005), a *lexical set* can be defined as the list of lemmas that frequently occur with a given target predicate in a given slot. For instance, the elements that can fill the object position of the verb *lire* ("to read") in sentence (4) are *article*, *libre* or *avis*, while the lexical set for the indirect object position associated with the verb *communiquer* ("to talk to") in sentence (5) are *ami*, *avocat* or *bébé*.

4. *Jean lisait un* {*article, libre, avis*}
   "Jean was reading a {article, book, advertisement}"
5. *Il a communiqué avec un* {*ami, avocat, bébé*}
   "He talked to a {friend, lawyer, baby}"

In *LexIt*, lexical sets are used to infer the semantic classes of the prototypical fillers of each semantic role of a SCF, by exploiting a variation of the algorithm proposed by Schulte im Walde (2006), that exploits WordNet supersenses (Fellbaum, 1998) as the reference inventory of semantic classes: ACT, ANIMAL, ARTIFACT, ATTRIBUTE, BODY, COGNITION, COMMUNICATION, EVENT, FEELING, FOOD, GROUP, LOCATION, MOTIVE, OBJECT, PERSON, PHENOMENON, PLANT, POSSESSION, PROCESS, QUANTITY, RELATION, SHAPE, STATE, SUBSTANCE, TIME.
In this way, the selectional preferences that can be inferred from the lexical sets in sentence (4) can be characterized as:

6. $[\text{PERSON}]_{\text{subj}}$
   lire
   $[\text{COMMUNICATION-ARTIFACT}]_{\text{obj}}$

If lexical sets describe the behavior of verbs as observed in a corpus, selectional preferences make an important generalization over the semantic properties of arguments.

3.3. A Resource on Italian Argument Structure

The database described by Lenci et al. (2012) has been built by applying the *LexIt* framework to the 'La Repubblica' (Baroni et al., 2004) corpus (ca. 331 millions tokens) and to a dump of the Italian section of Wikipedia (ca. 152 millions of tokens). The resulting dataset in the former setting encodes 3,873 verbs, 12,766 nouns and 5,559 adjectives, while the latter setting resulted in the characterization of 2,831 verbs and 11,056 nouns.
The *LexIt* extracted methodology has been evaluated by comparing the SCF frames available in three gold standard dictionaries for 100 test verbs against those automatically extracted from the 'La Repubblica' corpus, filtered by exploiting either a MLE-based threshold or a LMI-based threshold (see section 5). In the MLE-based setting, the authors reported 0.69-0.78 precision, 0.91-0.97 recall and 0.78-0.82 F-measure; while in the LMI-based setting the system obtained 0.77-0.82 precision, 0.92-0.96 recall and 0.84-0.85 F-measure.

4. Adapting the *LexIt* Framework to French

When compared with other state-of-the-art extraction models, the *LexIt* framework has a series of advantages, among which those that are crucial for our purposes are:

- the fact that the most salient frames are identified in a unsupervised manner. That is, it is not based on a pre-compiled list of valid SCFs, as it is the case for the *VALEX* model;

- the fact that this methodology can be applied to different parts of speech;
- the fact that it is an open and parameterizable framework, easily adaptable to novel languages or domains.

The acquisition framework consists of three modules: a *dependency extractor*, which extracts pattern for each target verbs; a *subcategorization frame identifier*, which filters patterns and extract lemmas associated to each argument position; and a *profile builder*, which finally builds the complete distributional profile. The process of adaptation to French affected only the first module.

4.1. Dependency Extractor

The goal of the first module is to analyze a dependency-parsed corpus in order to identify the occurrences of each predicate and to extract, for each occurrence with: the list of dependencies headed by the target predicate; the lexical elements filling each syntactic position. This process is carried out by an algorithm developed to filter and interpret the linguistic annotation available in the input. As a consequence, the design of this algorithm is strictly dependent on the properties of the linguistic annotation available in the corpus. Furthermore, we agree with those scholars (Preiss et al., 2007, inter alia) suggesting that the calibration of this module on the behavior of the specific parser has the effect of reducing the parser-specific bias in the input data.
In the original *LexIt* framework, data were extracted from the linguistic annotation realized by the Part-Of-Speech tagger described in Dell'Orletta (2009), together with the dependency parser DeSR (Attardi and Dell'Orletta, 2009). For this first release of *LexFr*, we tailored our extraction algorithm on the annotation provided by *Talismane*, an open-source suite of NLP tools proving a transition based statistical dependency parser (Urieli, 2013a; Urieli, 2013b). Accordingly, our first step has focused on the development of a set of specific pattern rules to:

- extract simple dependency relations such as subject, object, predicative complements, complements, finite and infinite clauses (and modifiers for verbs and nouns), as well as complex dependencies mediated by a preposition;
- handle problematic phenomena like the conversion of the passive diathesis, the identification of the antecedents of relative pronouns and the lemmatization of nominal predicates as *être heureux* ("to be happy") as a single lemma predicate-copula;
- extract other morphosyntactically relevant information, such as the presence of the reflexive pronoun *se*.

4.2. Subcategorization Frame Identifier

Data extracted in the first phase are processed in order to select relevant target predicates and fillers. By default, this is accomplished by resorting to a frequency threshold, but it is possible to exploit a combination of whitelists and blacklists to select only specific sets of lexical items.
The main goal of this step, however, is the identification of the argument structure licensed by each predicate occurrence. Such a process requires a list of allowed SCFs, sorted by frequency, that can be either selected by the user or automatically created by computing the frequency of all the possible slot combinations (e.g., `subj#obj#comp`$_{\text{à}}$, `subj#obj`, `subj#comp`$_{\text{à}}$, etc.) attested in the corpus and discarding those below a given frequency threshold.
By resorting to this list, our algorithm identifies the SCF licensed by each predicate in each sentence as the longest and most frequent unordered concatenation of argument slots. Notwithstanding its relative simplicity, this strategy turned to be significantly effective to limit the influence of both annotation errors (e.g., wrong syntactic parses) and marginally relevant dependency patterns, often due to a idiosyncratic sequences of adjuncts in a sentence.

4.3. Profiler

In the last steps, the system elaborates the distributional information filtered in the first two modules to build the final *distributional profiles*. From the output of the second module, this is obtained by:

- categorizing the fillers into WordNet supersenses by following the general methodology described by Resnik (1996). To extract the candidate synsets and general classes we resorted to the *Wordnet Libre du Français* lexicon (Sagot and Fišer, 2008) available in the *Open Multilingual WordNet* repository (Bond and Paik, 2012; Bond and Foster, 2013);
- aggregating the single co-occurrence for each information of interest (i.e., slots, SCFs, fillers, semantic classes), thus collecting, for each predicate of interest, its joint frequency with: 1. each SCF; 2. each slot (in isolation or in the context of each SCF) ; 3. each filler for given a slot (in isolation or in the context of each SCF) ; 4. each semantic class (in isolation or in the context of each SCF).
- calculating the strength of association to be loaded in the *distributional profiles*. Various weighting measures can be selected, among which relative frequency and common association measures (Evert, 2009).

5. A French Distributional Lexicon

Table 3 summarizes the lexical coverage of the current release of our French resource *LexFr* lexicon, including distributional knowledge extracted from *FrWaC*, a 90M tokens web corpus developed in the context of the *WaCKy* project (Baroni et al., 2009) and automatically annotated with the *Talismane* toolkit (Urieli, 2013a).
Following the design feature adopted in the Italian resource, we used Local Mutual Information (Evert, 2009, LMI) to weight the combinatorial properties of our target predicates with respect to a given context (e.g. a SCF, a slot, a filler...):

$$LMI(c_i, w_j) = f(c_i, w_j) * \log_2 \frac{p(c_i, w_j)}{p(c_i) * p(w_j)}$$

where $f(c_i, w_j)$ is the joint frequency of the predicate w_j with the context c_i, $p(c_i, w_j)$ is the joint probability of these entities, and $p(w_j)$ and $p(c_i)$ are the marginal probabilities

LexFr with *LexSchem* SCFs		*LexFr* with unsupervised SCFs	
SCF	LMI	SCF	LMI
subj#inf$_{de}$	1388.93	subj#inf$_{de}$	1390.15
subj#obj#comp$_{à}$	683.06	subj#obj#comp$_{à}$	683.06
subj#comp$_{à}$#inf$_{de}$	420.8	subj#comp$_{à}$#inf$_{de}$	420.8
subj#comp$_{à}$	248.9	subj#comp$_{à}$#fin$_{que}$	220.63
subj#si#inf$_{de}$	164.37	subj#si#inf$_{de}$	151.47
subj#comp$_{par}$	80.21	subj#comp$_{à}$	137.3
subj#obj#inf$_{de}$	53.40	subj#comp$_{par}$	80.13
subj#comp$_{à}$#comp$_{de}$	4.94	subj#obj#inf$_{de}$	53.7
subj#obj#comp$_{à}$#inf$_{de}$	4.79	subj#fin$_{que}$	43.93
subj#obj#comp$_{par}$	1.92	subj#si#fin$_{que}$	11.47

Table 2: syntactic profiles for the verb *promettre* ("to promise") built by exploiting two different sets of SCFs: an inventory extracted from *LexSchem* (left) vs. an automatically created list (right).

Part of Speech	no of lemmas	no of SCFs
verbs	2,493	99
nouns	7,939	99
adjectives	2,628	52

Table 3: distribution of target predicates and number of SCFs in *LexFr* (minimum frequency = 100)

of the predicate and of the context, respectively. LMI corresponds to the Pointwise Mutual Information (Church and Hanks, 1991, PMI) between the predicate and the context weighted by their joint frequency, and differs from PMI in avoiding the bias towards low-frequency events.

Table 2 reports two different sorted list of SCFs associated with the verb verb *promettre* ("to promise"): the left one is based on an inventory of 88 SCFs extracted from *LexSchem*; the right one is instead based on a list of 99 SCFs automatically extracted from *FrWaC*. By using LMI, we are able to see which verb-SCF pairs occur with a frequency that is higher that what could be expected if the verb and the frame were independent. For instance, while the verb *promettre* occurs with the zero-argument construction more times (frequency=815) than with the subj#inf$_{de}$ and the subj#obj#comp$_{à}$ frames (465 and 334 times, respectively), the low association score (LMI = -125.74 in the *LexSchem* setting) is a strong indication that the zero-argument construction is not relevant to describe the syntactic behavior of our target verb.

Moreover, the comparison between the results of the two settings reported in Table 2 seems to suggest is that the choice between the two SCFs lists does not dramatically change the syntactic representation extracted from the corpus: the lists of the top-associated SCFs are very similar in the two settings, and so are the association values of the SCFs, the only significant difference being the presence of the subj#comp$_{à}$#fin$_{que}$ SCF in the unsupervised setting. Needless to say, a more extended test of this conclusion is needed. As a design choice, in the release version of *LexFr* we opted for the totally unsupervised setting.

Table 4 depicts part of the semantic profile (lexical set and selectional preferences) for the complement introduced by à ("to") for the verb *promettre* within the frame subj#obj#comp$_{à}$. As stated by Lenci et al. (2012), this information has a twofold function: descriptive and predictive. Its descriptive role is fulfilled by the fillers ability to provide a sort of snapshot of the most representative words that co-occur with a predicate in a given slot. The characterization of the semantic classes of these fillers, on the other side, allows us to make predictions about possibly unseen fillers and to represent general semantic constrains on predicate slots.

5.1. Evaluation

In order to test the goodness of the subcategorization information available in *LexFr*, we evaluated the SCFs extracted by our system for a subset of 20 randomly sampled verbs (with frequency ≥ 400: see Appendix) against those attested in the other automatic extracted French lexicon, *LexSchem*. To compare the two lexicons, the *LexSchem* SCF format was converted into the *LexFr* format.

It is common practice to evaluate SCF extraction methods by filtering the output SCFs types with respect to some statistical scores, in order to filter out irrelevant frames (Korhonen, 2002). As an evaluation measure, we computed precision (the proportion of *LexFr* SCFs that are attested in the gold standard), recall (the proportion of *LexSchem* SCFs that have been extracted by our system) and F-measure (i.e. the harmonic mean of precision and recall). To rank and filter our SCFs, we resorted to the following scores:

- Maximum Likelihood Estimation (MLE), corresponding to the relative frequency of a SCF with a target verb:
$$f_{rel}(scf_i, v_j) = \frac{f(scf_i, v_j)}{f(v_j)}$$
where $f(scf_i, v_j)$ is the joint frequency of the verb v_j with the SCF scf_i, while $f(v_j)$ is the number of verb occurrences in the corpus;
- $LMI(scf_i, v_j)$: that is, the Local Mutual Information between the verb v_j and the SCF scf_i.

Precision, recall and F-measure were calculated at increasing thresholds of MLE and LMI. Figure 1 and Figure 2 plots

Lexical Set		Selectional Preferencies	
FILLER	LMI	CLASS	LMI
avenir (“future”)	264.60	PERSON	73.49
carrière (“career”)	76.76	TIME	23.97
disciple (“pupil”)	17.64	ARTIFACT	14.92
électeur (“voter”)	16.20	QUANTITY	3.07
ami (“friend”)	15.90	PLANT	1.13
affamè (“hungry man”)	8.85	EVENT	0.82
homme (“man”)	7.32	PROCESS	0.23

Table 4: semantic profile characterizing the complement introduced by *à* held by the verb *promettre* (“to promise”) within the frame `subj#obj#comp`$_{\text{à}}$.

F-measure values against increasing MLE and LMI thresholds, respectively.

In the MLE-based setting, the best scores were recorded with a relative frequency threshold around 0.15, where our system precision, recall and F-measure scores are around 0.74, 0.66 and 0.70, respectively. The LMI-based setting, on the other side, appears to be quite more complex, and the best scores (0.59, 0.60, 0.60) are typically obtained with a threshold between 100 and 200.

Overall, these results are consistent with those by Lenci et al. (2012), and point towards the effectiveness of such simple techniques in their ability to filter out possible noisy frames (Korhonen, 2002). Differently from the results of the *LexIt* evaluation, we couldn't find any advantage of recall over precision, and in some settings we even found evidence of the inverse pattern (e.g. 0.74 precision vs 0.66 recall in the best MLE setting). We interpret this phenomenon as the consequence of the different nature of the gold standards in the two evaluations: manually annotated in the case of *LexIt*, automatically extracted in the case of *LexFr*.

More puzzling, on the other side, is the advantage of MLE over LMI, especially when precision is involved. This is at odd with the results by Lenci et al. (2012) and can be accounted for in many ways: as a consequence of the nature of *LexSchem* or of the small sample size of the evaluation; as a suggestion of the inappropriateness of LMI and of association measures in general when SCF-like structures are involved. We leave such an issue open to future investigations, together with the evaluation of the other distributional information extracted with *LexFr*.

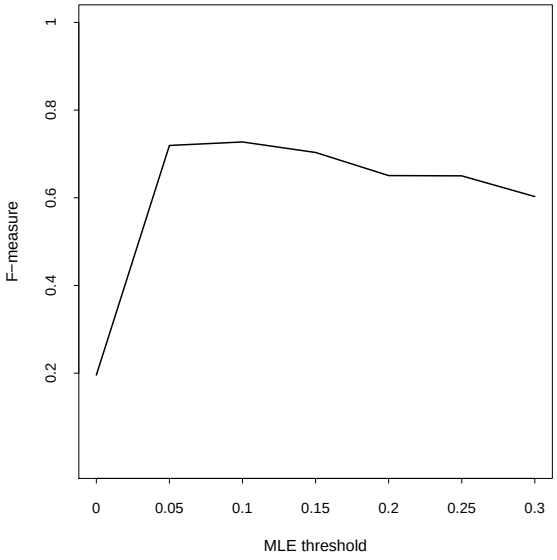

Figure 1: SCF F-measure and MLE threshold

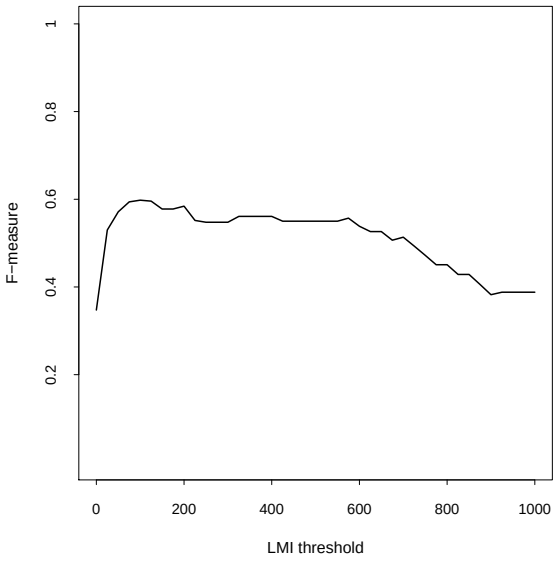

Figure 2: SCF F-measure and LMI threshold

6. Conclusion

The main purpose of our work was to show how the *LexIt* framework could be easily adapted to other languages in order to extract distributional profiles for verbs, nouns and adjectives. The case reported in these pages pertains to French, and the resulting combinatorial lexicon, *LexFr*, has been evaluated by comparing the information extracted for 20 test verbs against the relevant information available in a gold standard dictionary. The reported accuracy is in line with the state-of-the-art, thus supporting the crosslingual adaptability of the *LexIt* framework.

Ongoing work on this project includes:

- exploiting the *LexIt* framework to extract domain-relevant information;

- testing the possibility to apply this framework to other languages, including those more distant from romance ones;
- developing novel techniques to identify SCFs, weight their statistical significance, and characterize selectional preferences;
- enriching the resources with distribution information concerning MultiWord Expressions.

7. Acknowledgements

The authors are grateful to dr. Franck Sajous for providing the parsed version of the *FrWaC* corpus. This work received support from the CombiNet project (PRIN 2010-2011 *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, n. 20105B3HE8), funded by the Italian Ministry of Education, University and Research (MIUR).

8. Bibliographical References

Attardi, G. and Dell'Orletta, F. (2009). Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 261–264.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., and Mazzoleni, M. (2004). Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1771–1774.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bond, F. and Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1352–1362.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Bresnan, J. (1996). Lexicality and Argument Structure. In *Paris Syntax and Semantics Conference*.

Church, K. W. and Hanks, P. (1991). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, March.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*.

Dubois, J. and Dubois-Charlier, F. (1997). *Les Verbes Français*. Larousse-Bordas.

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

Evert, S. (2009). Corpora and Collocations. In A. Lüdeling et al., editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter.

Fellbaum, C. (1998). *WordNet - An Electronic Lexical Database*. The MIT Press.

Gross, M. (1975). *Méthodes en syntaxe: régime des constructions complétives*. Hermann.

Hanks, P. and Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue française de linguistique appliquée*, 10(2):63–82.

Korhonen, A., Krymolowski, Y., and Briscoe, T. (2006). A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the 5th Edition of the Language, Resources and Evaluation Conference (LREC 2006)*, pages 1015–1020.

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

Korhonen, A. (2009). Automatic Lexical Classification - Balancing between Machine Learning and Linguistics. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*, pages 19–28.

Kupść, A. and Abeillé, A. (2008). Growing TreeLex. In *Proceedings of the 9th Internation Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2008)*, pages 28–39.

Kupść, A. (2009). TreeLex Meets Adjectival Tables. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 203–207.

Lenci, A., Lapesa, G., and Bonansinga, G. (2012). LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3712–3718.

Levin, B. and Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge University Press.

Levin, B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press.

Light, M. and Greiff, W. (2002). Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26(3):269–281.

McCarthy, D. (2001). *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.

Messiant, C., Korhonen, A., and Poibeau, T. (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 533–538.

Messiant, C. (2008). A subcategorization acquisition system for French verbs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*, pages 55–60.

Preiss, J., Briscoe, T., and Korhonen, A. (2007). A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), pages 912–919.

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Roland, D. and Jurafsky, D. (2002). Verb sense and verb subcategorization probabilities. In S. Stevenson et al., editors, *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, pages 325–346. John Benjamins.

Sagot, B. and Danlos, L. (2012). Merging syntactic lexica: the case for French verbs. In *Proceedings of the LREC 2012 workshop Merging Language Resources*.

Sagot, B. and Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex 2008 Workshop at LREC*, pages 14–19.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*.

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

Schulte im Walde, S. (2009). The induction of verb frames and verb classes from corpora. In A. Lüdeling et al., editors, *Corpus Linguistics. An International Handbook*, chapter 61, pages 952–972. Mouton de Gruyter.

Urieli, A. (2013a). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Urieli, Assaf et Tanguy, L. (2013b). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions: études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence du Traitement Automatique du Langage Naturel (TALN 2013)*.

van den Eynde, K. and Blanche-Benveniste, C. (1978). Syntaxe et mécanismes descriptifs: présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32:3–27.

van den Eynde, K. and Mertens, P. (2010). Le dictionnaire de valence DICOVALENCE: manuel d'utilisation. Technical report, Université de Leuven.

Appendix: List of test verbs

abaisser	acheter	boire
composer	considérer	continuer
élaborer	équivaloir	orienter
pleurer	préserver	présider
qualifier	raser	remplir
réparer	repasser	retourner
réfugier	suspendre	