UDLex: Towards Cross-language Subcategorization Lexicons

Giulia Rambelli, Alessandro Lenci, Thierry Poibeau

UDLex: Towards Cross-language Subcategorization Lexicons

Giulia Rambelli and **Alessandro Lenci**
Computational Linguistics Laboratory
Department of Philology, Literature, and Linguistics
University of Pisa
Pisa, Italy
g.rambelli1@studenti.unipi.it
alessandro.lenci@unipi.it

Thierry Poibeau
LATTICE
CNRS, École normale supérieure and Université Sorbonne nouvelle
PSL Research University and USPC
Paris, France
thierry.poibeau@ens.fr

Abstract

This paper introduces *UDLex*, a computational framework for the automatic extraction of argument structures for several languages. By exploiting the versatility of the Universal Dependency annotation scheme, our system acquires subcategorization frames directly from a dependency parsed corpus, regardless of the input language. It thus uses a universal set of language-independent rules to detect verb dependencies in a sentence. In this paper we describe how the system has been developed by adapting the *LexIt* (Lenci et al., 2012) framework, originally designed to describe argument structures of Italian predicates. Practical issues that arose when building argument structure representations for typologically different languages will also be discussed.

1 Introduction

The argument structure of predicates is a key research area in Natural Language Processing (NLP), as verb valency has a decisive impact on sentence structure. Since including information about the syntactic-semantic realization of predicate arguments in a lexicon proved to benefit many NLP applications, e.g. recognition of textual entailment, information retrieval, machine translation and word-sense disambiguation (Korhonen, 2009), research in the (semi-)automatic acquisition of argument structure information from corpora has become widespread. Meanwhile, the last years have also witnessed a growing interest in multilingual studies and evaluation campaigns to test the quality and the robustness of parsing software.

By combining these two computational linguistic topics, our work is oriented towards the elaboration of a cross-language subcategorization lexicon, i.e. an automatically-built resource that encodes combinatorial properties of verbs at the syntax-semantics interface. This resource will in turn help the comparison of results among languages. In this paper, we describe the first steps into the realization of this resource, consisting in proposing a general framework to automatically derive verb subcategorization frames regardless of the specificities of the input language. For our purpose, we decided to exploit Universal Dependencies[1] (UD) annotations: UD is developed by the UD community with the final goal of creating a cross-linguistically consistent treebank annotation scheme for many languages (Nivre, 2015). The actual UD design combines the (universal) Stanford dependencies (de Marneffe and Manning, 2008; de Marneffe et al., 2014), the Google universal part-of-speech tags (UPOS) (Petrov et al., 2012) and the Interset interlingua for morpho-syntactic tag sets (Zeman and Resnik, 2008).

The aim of our project is twofold: on the one hand, we want to test if UD relations are sufficient to describe argument structure for some representative languages, and on the other hand we want to create a multilingual subcategorization lexicon to carry out a contrastive study regarding argument structures, i.e., the analysis of the syntactic realization patterns of verbs arguments across languages. For instance, we would like to know if synonymous predicates across languages occur with similar or different morpho-syntactic frames, or if the same valency frame in two languages is instantiated or not by similar constructions. Our aim is so to exploit UD treebanks to explore possible language universals concerning the relationship between form and meaning in argument structures. This work is the first step into building a unique database where all languages are aligned,

[1] www.universaldependencies.org

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 207-217,
Pisa, Italy, September 18-20 2017

in order to facilitate the comparison among lexica, using *FrameNet* (Fillmore, 1982; Fillmore, 1985) with links between verbs expressing similar semantic frames across different languages. A frame is a schematic representation of the situations that characterizes human experience, constituted by a group of participants in the situation (Frame Elements), and representing the possible syntactic realizations of the Frame Elements for every word (Fillmore and Atkins, 1992).

The paper is organized as follows: in section 2, we summarize related works on automatic lexical acquisition; in section 3, we describe the key characteristics of the *LexIt* framework and we then focus on the adaptation of the original module to the UD annotation scheme (section 4). We then describe the resulting lexica for English, Italian, French, German and Finnish. We conclude with a general discussion about argument representation (section 5). Ongoing work will be discussed in section 6.

2 Previous work

Automatic lexical acquisition, that is the research area that develops methodologies to automatically build large-scale, wide coverage lexical resources, is constantly growing and lots of resources have been built for several languages. Among the several kinds of information that can be acquired from a corpus, it is worth mentioning the intrinsic relation between the semantics of a predicate and the morpho-syntactic realization of its arguments, embracing the theoretical assumption described by (Levin, 1993; Bresnan, 1996; Roland and Jurafsky, 2002; Levin and Rappaport-Hovav, 2005).

In the last two decades, automatic methods have been developed for the identification of verb subcategorization frames (SCFs) (Korhonen, 2002; Messiant et al., 2010; Schulte im Walde, 2009), selectional preferences (Resnik, 1996; Light and Greiff, 2002; Erk et al., 2010) and diathesis alternation (McCarthy, 2001). The approach consists in automatically infering subcategorization frames directly from the corpus, with or without a predefined list of possible frames. The literature reports a large number of automatically built subcategorization lexica, among which *VALEX* for English verbs (Korhonen et al., 2006), *LexSchem* (Messiant et al., 2008) and *LexFr* (Rambelli et al., 2016) for French verbs, *LexIt* for Italian verbs, nouns and adjectives (Lenci et al., 2012). SCFs acquisition has been investigated also for languages such as Chinese (Han et al., 2004) and Japanese (Marchal, 2015). These resources have been of particular interest to classify verbs on the basis of their syntactic and semantic properties, producing several taxonomies comparable to *VerbNet* (Kipper-Schuler, 2005).

Despite the importance of these resources, existing lexica only focus on a single language with a specific syntactic frame representation, strongly dependent on the corpus used for acquisition. Few studies tried to automatically build multilingual SCFs lexica. To the best of our knowledge, there have been few experiments in multilingual verb lexicon with syntactic and semantic information, mostly establishing multilingual links manually (Civit et al., 2005; Hellan et al., 2014).

3 The *LexIt* Framework

LexIt (Lenci et al., 2012) is a computational framework whose aim is to automatically extract distributional information about the argument structure of predicates. It was originally developed to extract information on Italian verbs, nouns and adjectives from “La Repubblica” (Baroni et al., 2004) corpus (ca. 331 millions tokens) and from a “dump” of the Italian section of Wikipedia (ca. 152 millions of tokens). The database resulting from this previous work is freely browsable.[2] The whole framework aims at processing linguistic information from a dependency-parsed corpus and then storing the results into a database where each predicate is associated with a distributional profile, i.e. a data structure that combines several statistical information about the combinatorial behaviour of the lemma. This profile is articulated into:

1. a *syntactic profile*, specifying the syntactic arguments (a.k.a. syntactic *slots*: e.g. subject, complements, modifiers, etc.) and the subcategorization frames (SCFs) associated with the predicate;
2. a *semantic profile*, composed of:
   - the *lexical set* of the most typical lexical items that occur in each syntactic slots;
   - the *semantic classes* characterizing the selectional preferences of the different syntactic slots.

[2] http://lexit.fileli.unipi.it/

This framework was designed to be open and adaptable to novel languages and domains. For example, the most salient frames can be identified directly from corpora in an unsupervised manner, without the need to provide a pre-compiled list of valid SCFs (contrary to what was done for the VALEX model for example). Besides, there is no formal distinction between arguments and adjuncts: a SCF is represented as an unordered pattern of syntactic dependencies whose combination is strongly associated to the target predicate. But the key aspect is that the system consists of a pipeline of three modules:

Dependency extractor The first module extracts the syntactic dependencies of each predicate in a sentence along with the lexical elements realized in the slots. The inventory of slots for verbs comprehends subject (`subj`), object (`obj`), complements (`comp*`), finite clauses (`fin*`) and infinitives (`inf*`), including the presence of the reflexive pronoun (`se`) and predicative complements (`cpred`). The design of the algorithm is strictly dependent on the output of a specific parser.

SCF Identifier The main goal of this step is to identify SCFs licensed by each verb in a sentence using filtering techniques to remove possible noisy frames. Given a list of allowed SCFs, our algorithm identifies the SCF licensed by each predicate in each sentence as the longest and most frequent unordered concatenation of argument slots. The resulting frames are represented as a list of syntactic slots concatenated with the symbol "#". For instance, a subject-object transitive SCF is marked as `subj#obj`.

Profiler Finally, the system categorizes lexical elements into WordNet (Fellbaum, 1998) supersenses and compute selectional preferences by following the methodology described by Resnik (1996). The module builds the final profiles by computing for each predicate its joint frequency and strength of association with each SCF, each slot, each lexical element for a given slot (in isolation or in each SCF) and semantic class (in isolation or in each SCF).

The final *LexIt* dataset encodes 3,873 verbs, 12,766 nouns and 5,559 adjectives for "La Repubblica" corpus and 2,831 verbs and 11,056 nouns for Wikipedia dump. The resulting syntactic information has been evaluated by comparing the SCF frames available in three gold standard dictionaries against those automatically extracted from the "La Repubblica" corpus, filtered by exploiting either a MLE-based threshold or a LMI-based threshold. 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.

The system adaptability was also tested by using different existing modules for French. The result was the *LexFr* lexicon (Rambelli et al., 2016), representing information for 2,493 verbs, 7,939 nouns and 2,628 adjectives extracted from *FrWaC* web corpus of 90M token (Baroni et al., 2009). The evaluation of the automatically acquired frames against a gold standard dictionary was in line with the state-of-the-art (0.74 precision, 0.66 recall and 0.70 F-measure), thus supporting the cross-lingual adaptability of the *LexIt* framework.

4 UDLex: Adapting the *LexIt* Framework to UD

As said above, the dependency extractor is the only module of the *LexIt* framework to be strictly dependent on the annotation scheme of the input corpus. Therefore, a set of rules must be developed each time the system has to process a new language or a corpus with a different annotation scheme. To overcome this limitation, we decided to adapt the extractor algorithm to the Universal Dependency annotation scheme, a cross-linguistically consistent grammatical annotation. We also focused on some specific linguistic phenomena which vary from language to language and for this reason are treated in a specific way depending on the reference theoretical framework.

4.1 Universal Dependencies

As Manning (2015) states, the UD scheme was designed to optimize subtle trade-off between a satisfactory analysis on linguistic grounds and an annotation scheme that can be automatically applied to several languages with good accuracy. UD is not proposed as a linguistic theory, but rather as a good compromises in the interest of practical NLP applications, i.e., multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective (Nivre, 2015). Therefore, the representations adopted by UD are oriented towards surface syntax with a simple, lexically shallow approach that primarily focuses on transparently encoding predicate-argument structure.

The latest version 2.0 uses a more consistent and efficient annotation, even if UD teams still work on language-specific issues (there are still lots of inconsistencies in the migration from UD v1 and UD v2, for example regarding reflexive pronouns). The last release of UD treebanks covers 45 different languages. For what concerns syntactic relations, UD v2 contains 37 universal grammatical relations that re-arrange previous dependencies based on the *core-oblique* distinction (for more details, see (Thompson, 1997)). As stated in UD guidelines, this distribution is grounded on the assumption that all languages have some prototypical way of encoding the arguments of intransitive and transitive verbs, often referred to as S (for the subject of an intransitive verb), A (for the subject/agent of a transitive verb) and O or P (for the object/patient of a transitive verb). Each language has its own way to establish what is the prototypical encoding: it often involves some combination of case-marking (nominative-accusative or ergative-absolutive) and/or indexing on the verb (agreement) and/or linear position in the clause (typically relative to the verb). We can add to this the possibility to undergo certain grammatical transformations, such as relativization and passivization. In UD, the notion of core argument (nsubj, iobj, obj plus argument clauses) is reserved to those dependents of the verb that exhibit all or most of this prototypical encoding.

Accordingly, all other dependents of the verb are oblique, a fuzzy concept which entails different things for different languages. For example, in English it means having a prepositional marker and/or occurring in a different position relative to the verb than core arguments. For case languages, obliques may either be accompanied by adpositions or occur with cases that are not prototypical for core arguments (often referred to as oblique cases). Exactly which cases are regarded as oblique can again vary between languages, and typical borderline cases are dative, partitive and (less commonly) genitive[3]. Note also that a specific linguistic property, such as the presence of an adpositional marker, cannot be considered as a universally valid criterion for obliqueness. The core-oblique distinction should not correspond to argument-adjunct distinction. In a language like Italian or French, for example, prepositions are used in the prototypical encoding of indirect objects and prepositional complements can occur as arguments into a subcategorization frame.

[3]And of course, each language uses this terminology differently. We are well aware that a Finnish genitive has very little to do with a Latin genitive, for example.

4.2 Selected phenomena tackled by UDLex

4.2.1 Indirect object

In the UD scheme, the core argument *iobj* identifies a noun phrase that is generally the indirect object of a verb. In German and in languages distinguishing morphological cases, the indirect object is often marked by the dative case (even if it may take other forms as well). For these languages, we decided to include into the list of argument slots a new label `iobj`. So, sentences in (4) refers to a unique frame `subj#obj#iobj`). As English have also a double object construction, its frame list will admit both a `subj#obj#iobj` e `subj#obj#comp`$_{to}$ (examples in (1)). However, in Italian and French this relation only appears when the indirect object is a clitic pronoun, while if the indirect object is realized as a prepositional phrase it is marked with `obl` relation. In this perspective, sentences in (2) should be both represented with frame `subj#obj#comp`$_{a}$ and sentences in (3) with `subj#obj#comp`$_{à}$ slots, to avoid double object construction for these two languages.

(1) a. *The woman gives <u>him</u> an apple.*
    b. *The woman gives an apple <u>to the child</u>.*

(2) a. *La donna <u>gli</u> dà una mela.*
    b. *La donna dà una mela <u>al bambino</u>.*

(3) a. *La femme <u>lui</u> donne une pomme.*
    b. *La femme donne une pomme <u>à l'enfant</u>.*

(4) a. *Die Frau gibt <u>ihm</u> einen Apfel.*
    b. *Die Frau gibt <u>dem Kind</u> einen Apfel .*

4.2.2 Reflexive pronoun

The UD has a specific morphological feature *Reflex* that tells whether a given word is reflexive, i.e. refers to the subject of its clause. However, not all languages that have a reflexive pronoun use this label, preferring more elaborated kinds of annotation. For example, the team developing the Italian UD Treebank did not choose to include into the feature list this specific label, since this information does not seem to add relevant information for training a syntactic parser, and it is quite redundant with the presence of the language-specific label *Clitic*.

For Italian, we designed a simple rule that identifies into a sentence all pronouns that are 1) clitics (with the morphological feature `Clitic=Yes` and 2) the objects of verbs (`obj` relation). We also use a whitelist of admitted pronouns forms to avoid clitics that are real object of the verb.

(5) a. *Maria si lava.* "Mary washes herself".
b. *Maria li lava.* "Mary washes them".

In sentence (5), verb *lavare* ("wash") occurs with two clitic pronouns that are marked with the same label `obj`. However, the verb is reflexive only in (a) (`subj#si#0`, while it has the transitive frame `subj#obj` in (b). The algorithm detects the two forms by verifying that the form of the pronoun is included in the whitelist and that the verb and the pronoun agree in number and person. The Italian treebank still has lots of inconsistent annotations regarding the possible values of a clitic, e.g. the dependency `expl` that marks the impersonal form of a verb is sometimes used to label the reflexive pronouns.

French also uses this label in a different way, to identify the combination of the personal pronouns with the adjective "même/s" to emphasize on the person ("myself, yourself..."), while the reflexive pronoun is detected using the dependency relation *expl*. The expletive relation can be used for reflexive pronouns attached to inherently reflexive verbs, i.e. verbs that cannot occur without the reflexive pronoun (see Figure 1).

We have to clarify that actually the nature of these clitics is underspecified, so we do not distinguish among verbs which have lexicalized pronoun (e.g. *s'amuser* "to have fun"), verbs which alternate reflexive form with a transitive one (e.g. *se raser* and *raser* "to shave (one self)"), and verbs whose reflexive form expresses a reciprocal action between more than one person, (e.g. *s'aimer* "to love each other" or *se parler* "to talk to each other").

4.2.3 Passive voice

Our system takes into account a traditional argument syntactic alternation: the relation between active sentence and its passive counterpart. Following Chomsky (1957; 1965), the two forms of verbs actually rely on the identical subcategorization frame and share the same selectional preferences (in the so called *underlying semantic structure*), but they differ in their syntactic derivation (or *surface structure*). Given this assumption, our system tries to reduce the two forms into a single SCF entry, converting the subject of passive sentences into the verb object and the agent complement into the subject. Concerning languages that have a grammaticalized passive transformation (among all English, Italian, French, German), the subject of this passive sentences is labelled with the subtype *nsubj:pass*. More complex is inferring the subject of the active form from a passive sentences: for example, in Italian this is generally conveyed by the prepositional phrase introduced by *da* ("by"), as illustrated in figure 2. In this case, the algorithm identifies the verb *provocare*("to cause") and extracts the frame `subj#obj` instead of `subj#comp`$_{da}$.

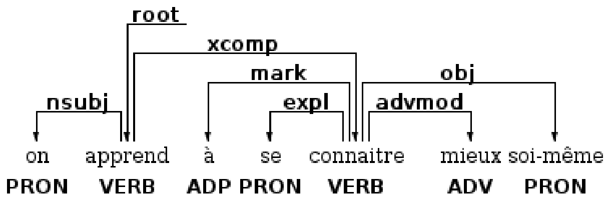

Figure 1: A French sentence with the reflexive pronoun ("We learn to know ourselves better").

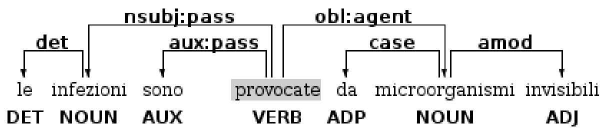

Figure 2: An Italian example of a passive sentence ("The infections are caused by invisible microorganisms").

However, the preposition *da* can express other complements, e.g. a locative or a temporal ones. In case the algorithm does not succeed in extracting the correct dependency of the verb, a subject slot with empty lexical is added to the resulting frame.

Note that the Finnish passive works quite differently and cannot be directly connected to an active form.

4.2.4 Co-reference in relative clauses

Our framework does not only detect the type of arguments of a given verb, but also store the lexical element in each slot. In order to store as many information as possible, it is useful to detect reference chains and try to re-annotate each pronoun with the appropriate antecedents. We consider in particular the case of relative pronoun. The UD created a specific relation *acl:relcl* for identifying the lexical antecedent of a relative clause. This label is used in 17 languages: Chinese, Danish, English, Estonian, Finnish, French, Greek, Hebrew, Hindi, Irish, Italian, Norwegian, Persian, Portuguese, Russian, Spanish, Swedish.

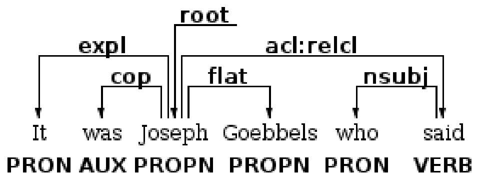

Figure 3: An example of relative clause annotation in English.

4.3 Resulting resources

The final system, *UDLex*, was run to extract syntactic frames and its lexical realization from Universal Dependencies 2.0 treebanks. As the corpora were released for the CoNLL 2017 shared task[4], we performed our experiments on available training sets. As a starting point, we tested *UDLex* on four languages: English, Italian, French and Finnish. Table 2 summarizes the characteristics of the input corpora.

	Tokens	Predicates	Lexical elements
English	229753	364	914
Italian	356912	481	1448
French	483781	543	1602
Finnish	181138	419	765

Table 1: Statistics in selected UD treebanks.

The resulting lexica mostly preserve the distributional profile format exploited in *LexIt* and *LexFr*. A verb syntactic profiles lists all the SCFs sorted by their salience, while the lexical set returns all the lexemes occurring in each slot of a SCFs. To identify prototypical or salient contexts of verbs (e.g. a SCF, a slot, a lexical realization of an argument), the system uses Local Mutual Information (Evert, 2009, LMI). In general, for a target word w_j and a context c_i, LMI is computed as follows:

[4]http://universaldependencies.org/conll17/

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

LMI is an association measure which corresponds to the verb-SCF joint frequency $f(c_i, w_j)$ weighted with Pointwise Mutual Information (PMI) between the v_j and the SCF scf_i. PMI quantifies the discrepancy between the probability $p(c_i, w_j)$ of verb-SCF coincidence and the probability $p(c_i)$ and $p(w_j)$ of their individual distributions, assuming independence. Unlike PMI, LMI reduces the risk of overestimating the significance of low-frequency events.

A slight difference compared to *LexIt* regards the presence of `iobj` label among admitted syntactic slots (see Table 2). This argument was included for those languages that need to mark the indirect object (section 4.2.1).

Label	Argument Slot
0	zero argument construction
subj	subject
si	reflexive pronoun
cpred	predicative complement
obj	direct object
iobj	indirect object
comp_*	prepositional phrases
fin_*	finite clauses
inf_*	infinitive clauses

Table 2: SCF argument slots.

Tables 3a–3c report the SCFs associated to the English verb *play* and its translation for Italian (*giocare*) and French (*jouer*). As the number of occurrences in the corpora is quite low (50, 58 and 141 respectively), there are very few really associated frames, while most of them occurs once with it the target predicate. However, it is possible to see some syntactic correspondences among the three tables, e.g. the presence of locative complement in several frames.

Table 4 instead lists extracted lexical items that occur as objects of target predicates. The English and French lexemes can be connected to three different semantic field: competition (*chess* in English vs *match*, *finale* in French), cause noise/music (*song* vs *chanson*) and perform a role (*role*, *part*, *movie* vs *rôle*, *personnage*). However, Italian verb *giocare* is not polysemic, in fact lexemes occurring in its context all refer to the com-

SCF	LMI
subj#obj#comp$_{in}$	14.10
subj#obj	9.56
subj#0	5.54
subj#comp$_{in}$#comp$_{with}$	3.13
subj#comp$_{with}$	1.80
subj#comp$_{in}$	

(a)

SCF	LMI
subj#comp$_{con}$	24.03
subj#comp$_{in}$	15.84
subj#comp$_{a}$	4.40
subj#comp$_{contro}$	4.29
subj#comp$_{per}$	3.38
subj#obj#comp$_{con}$	0.53

(b)

SCF	LMI
subj#obj#comp$_{dans}$	22.46
subj#obj#comp$_{avec}$	18.38
subj#comp$_{avec}$	17.74
subj#comp$_{dans}$	17.35
subj#comp$_{pour}$	16.81
subj#0	-13.77

(c)

Table 3: Syntactic profile of the verb *play*, *giocare* and *jouer*.

petition field (*ruolo* has to be intended as the role into a team).

A major limitation of this first experiment was the small dimension of existing treebanks. By filtering infrequent lemmas we obtained a narrow group of verbs, and the relative frequencies and association measures between a target verb and its SCFs are really lower, as shown in Tables 3a–3c. Moreover, the lexical sets consist of very few lexical item with a very low joint frequency.

English	Italian	French
role (86.8)	*partita* (78.7)	*rôle* (238.4)
chess (16.3)	*ruolo* (11.9)	*match* (58.1)
part (9.5)	*incontro* (6.9)	*personnage* (17.8)
song (6.6)	*gioco* (6.6)	*morceau* (11.8)
couple (5.9)		*chanson* (8.8)
movie (5.9)		*performance* (6.0)
version (5.4)		*finale* (4.1)

Table 4: Lexical sets of the object of *to play*, *giocare* and *jouer*. Between parentheses, the LMI values between each verb and the lexical filler.

4.3.1 Evaluation

The standard methodology for testing the accuracy of an automatically acquired subcategorization lexicon is to evaluate extracted SCFs against a manual annotated gold standard (Preiss et al., 2007). Although this approach may not be ideal (Poibeau and Messiant, 2008) in our case as we work with small corpora (so a dictionary may include a significant number of SCFs not attested in our data), it can provide a useful starting point.

For our purposes, the gold standard is represented by the valence patterns extracted from three manually-built lexical resources:

- *Valency Patterns Leipzig* (ValPaL) – an online database[5] that stores valency information for a small sample of verbs of 36 different languages, including English (Goddard, 2013) and Italian (Cennamo and Fabrizio, 2013). The aim of the project is to carry a cross-linguistic study of valency classes, choosing verbs that have the same meanings and encoding the valency information in a standard way.

- *Dicovalence* (Mertens, 2010) – a valency lexicon containing information for more than 3,700 French verbs. It is based on the pronominal approach (Eynde and Mertens, 2003), a linguistic theory that treats pronouns as semantic primitives due to the purely linguistic nature and a 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.

For each language, we selected the most frequent 20 verbs among those attested in both the gold standards and in the resulting lexicons. There are many differences in the way valence patterns are represented in gold standard and in *UDLex*, so checking which extracted frames also appear in the lexical resources is not a straightforward operation. Accordingly, we manually verified for each SCF whether it was attested in the gold standard or not. For example, ValPaL and *Dicovalence* use a general label for locative complements, with no information about the type of preposition involved, while *UDLex* considers all prepositions heading a slot as a distinctive feature for frames. In these cases, we regarded the extracted frames as correct, if the gold standard contains a frame with an acceptable prepositional phrase looking at the exam-

[5] `http://valpal.info`

ple sentences in the lexical resources (if available) or at corpus examples.

The standard practice to evaluate automatically-acquired SCFs is to filter frames with respect to some statistical score so as to exclude "noisy" frames caused by tagging or parsing errors. In particular, only SCFs with a score above a certain threshold are evaluated. We followed the same procedure resorting to Maximum Likelihood Estimation (Korhonen, 2002), that corresponds to the relative frequency of a scf_i with a verb v_j and it is calculated as follows:

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

We then computed precision (the proportion of extracted SCFs that are attested in the gold standard), recall (the proportion of gold SCFs that have been extracted by our system) and F-measure (i.e., the harmonic mean of precision and recall) over the three gold-standards for increasing thresholds of MLE in order to reach the best scores (Lenci et al., 2012).

Results are generally a bit lower than the state-of-the-art (see Table 5). For the three resources we obtained very high recall but low precision. The precision score is mostly affected by the fact that in *UDLex* our approach do not consider the argument/adjunct distinction, as it extracts all SCFs in an unsupervised way. On the contrary, the three gold standard resources (in particular ValPaL) code only core verb argument, ignoring possible adjuncts or circumstantial slots that could be meaningful in the description of the frame verb. This also explains why recall is higher than precision in all settings. To better understand the differences between the gold standard and the lexicons, we then performed a manual analysis (Poibeau, 2011).

	Precision	Recall	F-measure
En_ValPaL	0.49	0.62	0.55
Dicovalence	0.37	0.63	0.47
It_ValPaL	0.55	0.51	0.53

Table 5: Top scores with MLE thresholds.

UDLex has the best performance for English, because ValPaL encodes a very small set of possible SCFs (only 21 distinct and very basic frames can be extracted from the resource). All ValPaL frames are attested in our resource, but our system extracts a large number of other frames. For instance, *to call* is associated with only one frame in ValPal `subj#cpred#obj`, while 17 SCFs can be found in our lexicon, most of them being without doubt relevant like `subj#comp_for` (*I called for assistance*), `subj#obj` (*I called the hotel*), etc.

Another example is provided by the Italian reflexive pronoun *si*. ValPal encodes very fine-grained distinctions between different uses of *si*, such as true reflexive constructions, impersonal uses, pronominal intransitives, etc. Capturing these differences goes well beyond the expressive capability of our lexicon. As a matter of fact, for each languages our approach only distinguishes verb frames containing a reflexive pronoun (e.g.,`subj#si#0`), from those not containing any (e.g.,`subj#0`). Consistently, we decided to not consider more fined-grained distinctions in the present evaluation.

Among all languages, French obtains the worst results. *Dicovalence* is very different from ValPaL since it is based on a more fined-grained representation, leading to a number of 386 distinct subcategorization frames. For example, in *Dicovalence* there is a distinction between the verb *appeler* (to call) and the construction *en appeler*, that has the specific meaning "to appeal" (cf. *J'en appelle à votre bonté pour lui donner une deuxième chance*. "I appeal to your kindness to give him a second chance"). Obviously, this kind of information is difficult to automatically detect, and our resource does not contain this construction (although it is also questionable whether these are really two different, unrelated word senses).

5 Perspectives

The previous section introduced the distributional profiles resulting of the application of *UDLex* to English, Italian and French, i.e. closely related languages from a typological point of view. However we still have to further investigate whether the actual syntactic frame representation is sufficient for all kinds of languages , or if we should take into account additional morpho-syntactic phenomena when dealing with other, typologically-different, languages.

We need in particular to have a closer look at non Indo-European languages. In order to do this, we chose as a starting point to test our framework on Finnish, which is characterized by several interesting linguistic phenomena such as, inter alia, "differential object marking", which means that the object of a given verb may be marked by different cases (esp. nominative, genitive, accusative or partitive), depending on the verb, the noun and the overall meaning one wants to convey (for a more detailed description, see Karlsson (2008)). Chaminade and Poibeau (2017) studied this phenomenon by automatically extracting Finnish predicative structures from corpora. They then categorized verbs into three categories: verbs subcategorizing exclusively the partitive case, verbs subcategorizing exclusively the accusative/genitive case and verbs subcategorizing both cases.

(6) *Poika lukee kirjaa.* "the boy is reading a/the book" (as opposed to *Poika lukee kirjan.*, where *kirjan* is the genitive form and the whole sentence is resultative).

Sentence (6) is a simple example of a sentence with a transitive verb and a partitive complement. Thanks to UD annotation, our actual system induces a frame `subj#obj`, where the subject is *poika* and the object is *kirjaa*. However, an alternative possible representation of the frame would be `subj#obj+partitive`, including information about the case of the object. In this example, the partitive case means that the action is not completed, but the same sentence with `subj#obj+genitive` (*kirjan*) would also be entirely valid, with emphasis on the finiteness and totality of the clarification. As this distinction refers to the verbal aspect, we need to decide whether we want to include the representation of object cases or not.

Other features should be studied in greater detail. For example, Finnish has a so-called passive form (*Luetaan kirja/kirjaa*), but it can hardly be analyzed as being the transformation of a corresponding active form. The Finnish passive is available only for the 3rd person singular, and in fact corresponds to an active form with an unspecified subject. Moreover this form is used in various contexts, and can be either an injunction to do something ("let's read a book!") or can just be used instead of the 1st person plural in speech and dialogue. All this is of course known from traditional grammars but a general framework like UD may help us reconsider terminological issues and thus clarify the linguistic analysis of frequent word forms.

Passive is not the only example one can give when considering a language as different from Indo-European as Finnish. One should also consider null subjects used for "generic sentences expressing a general truth or law or state of affairs" (Karlsson, 2008) (Karlsson gives the following examples: *Usein kuulee, että...* "One often hears that..." or *Siellä saa hyvää kahvia.* "One gets good coffee there"). One should also consider sentences expressing an obligation, where the person affected is expressed through a genitive (*Miesten on pakko poistua.* "The men have to leave") or other sentences expressing a transformation (*Hänestä tuli lääkäri* "He has become a doctor", where the source of the transformation is expressed through a special case called elative). All this should be taken into account when processing Finnish corpora and it is not fully clear yet what should be taken into consideration during the analysis (as opposed to language idiosyncrasies that should be left apart), what is part of the dictionary (as opposed to a more general syntactic level) and how to deal with all this in a multilingual framework.

6 Conclusion

In this paper, we have proposed a general framework making it possible to build SCF lexicons for all the languages with a UD annotated corpus. The main purpose of our work was to understand how the UD annotation scheme represents information about verb dependencies in different languages. Our preliminary results show that our main algorithm is able to detect essential information about subcategorization frames for every languages exploiting general UD relations. Furthermore, the modularity of the framework makes it possible to process different language, taking ionto account language specificities with minimal changes.

Ongoing work includes the development of strategies to link lexica for different languages using the notion of "shared semantic frames". Our approach is based on a contextualized distributional analysis of argument structures, that is, we plan to exploit the distribution of lexical items in the different SCFs of a given verb to cluster verb senses, as already explored by Rumshisky (2008). Furthermore, we plan to link SCFs of verbs from different languages by combining bilingual dictionaries with information about the semantics of their respective arguments.

Finally, we are considering a practical evaluation through the integration of this resource into specific natural language applications. The results presented in this study can be seen as a first step in creating a multilingual subcategorization lexicon based on a pure distributional approach rather than a manually-built resource.

Acknowledgments

This work was partially funded by the ANR ERA-NET ATLANTIS project. Giulia Rambelli has benefited from an Erasmus grant while visiting the Lattice Lab.

References

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 the 4th International Conference on Language Resources and Evaluation (LREC'04)*:1771–1774.

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

Joan Bresnan. 1996. Lexicality and Argument Structure. In *Paris Syntax and Semantics Conference*.

Michela Cennamo and Claudia Fabrizio. 2013. Italian Valency Patterns. In I. Hartmann, M. Haspelmath and B. Taylor (Eds.), *Valency Patterns Leipzig.* Max Planck Institute for Evolutionary Anthropology, Leipzig.

Guersande Chaminade and Thierry Poibeau. 2017. Preliminary Experiments in the Extraction of Predicative Structures from a Large Finnish Corpus. In *Proceedings of the Workshop 3rd International Workshop for Computational Linguistics of Uralic Language*:37–55.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA.

Montserrat Civit, Joan Castelvi, Roser Morante, Antoni Oliver, and Joan Aparicio. 2005. 4LEX: a Multilingual Lexical Resource. In *Cross-Language Knowledge Induction Workshop*:39–45.

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

Stefan Evert. 2009. Corpora and Collocations. In A. Lüdeling et M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin.

Karel van den Eynde and Piet Mertens. 2003. La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104.

Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database.* The MIT Press, Cambridge, MA.

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 (Eds.), *Frames, fields and contrasts*:75–102. Lawrence Erlbaum Associates, Hillsdale, NJ.

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

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

Cliff Goddard. 2013. English Valency Patterns. In I. Hartmann, M. Haspelmath and B. Taylor (Eds.), *Valency Patterns Leipzig.* Max Planck Institute for Evolutionary Anthropology, Leipzig.

Xiwu Han, Tiejun Zhao, Haoliang Qi, and Hao Yu. 2004. Subcategorization acquisition and evaluation for Chinese verbs. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*.

Lars Hellan, Dorothee Beermann, Tore Bruland, Mary Esther Kropp Dakubu, and Montserrat Marimon. 2014. MultiVal towards a multilingual valence lexicon. In *Proceedings of the 9th Edition of the Language, Resources and Evaluation Conference (LREC'14)*:2478–2485.

Fred Karlsson. 2008. *Finnish: An Essential Grammar*. 2nd edition. Routledge Essential Grammars, London.

Karin KipperSchuler. 2005. VerbNet: A Broadcoverage, Comprehensive Verb Lexicon. PhD thesis, University of Pennsylvania, Philadelphia, PA. .

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the 5th Edition of the Language, Resources and Evaluation Conference (LREC'06)*:1015–1020.

Anna Korhonen. 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)*:19–28.

Anna Korhonen. 2002. *Subcategorization acquisition*. PhD thesis, University of Cambridge.

Alessandro Lenci, Gabriella Lapesa, and Giulia Bonansinga. 2012. LexIt : A Computational Resource on Italian Argument Structure. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*:3712–3718.

Beth Levin and Malka Rappaport-Hovav. 2005. *Argument Realization*. Cambridge University Press, Cambridge, UK.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, IL.

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

Christopher D. Manning. 2015. The case for universal dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*:1.

Pierre Marchal. 2015. *Acquisition de schmas prdicatifs verbaux en japonais*. PhD Thesis, INaLCO.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*:4585–4592.

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

Piet Mertens. 2010. Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE. Conversion vers un format utilisable en TAL. In *Actes TALN 2010*.

Cédric Messiant, Thierry Poibeau, and Anna Korhonen. 2008. Lexschem: a large sub-categorization lexicon for French verbs. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'08)*:142–147.

Cédric Messiant, Kata Gábor, and Thierry Poibeau. 2010. Lexical acquisition from corpora: the case of subcategorization frames in French. *Traitement Automatique des Langues*, 51(1):65–96.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In: Alexander Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2015*:3–16. Springer, Cham.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8ht International Conference on Language Resources and Evaluation (LREC'12)*:2089–2096.

Thierry Poibeau and Cdric Messiant. 2008. Do we still need gold standard for evaluation ? In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*.

Thierry Poibeau. 2011. *Traitement automatique du contenu textuel*. Lavoisier. Paris.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics (ACL'07)*:912–918.

Giulia Rambelli, Gianluca E. Lebani, Alessandro Lenci and Laurent Prévot. 2016. LexFr: adapting the LexIt framework to build a corpus-based French subcategorization lexicon. In *Proceedings of the 10th Edition of the Language, Resources and Evaluation Conference (LREC'16)*:930–937.

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

Douglas Roland and Daniel Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In Paola Merlo and Suzanne Stevenson (Eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*:325–346. John Benjamins, Amsterdam.

Anna Rumshisky. 2008. Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. In *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica.*

Sabine Schulte im Walde. 2002. A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC'02)*:1351–1357.

Sabine Schulte im Walde. 2009. The induction of verb frames and verb classes from corpora. In A. Lüdeling et M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, chapter 61. Mouton de Gruyter, Berlin.

Sandra A. Thompson. 1997. Discourse Motivations for the Core-Oblique Distinction as a Language Universal. In Akio Kamio (Ed.), *Directions in Functional Linguistics*:59–82. Benjamins, Amsterdam.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages*.