Flexible Acquisition of Verb Subcategorization Frames in Italian

Tommaso Caselli, Francesco Rubino, Francesca Frontini, Irene Russo and Valeria Quochi
ILC-CNR
Via G. Moruzzi, 1 56124 Pisa
E-mail: {tommaso.caselli}{francesca.frontini}{valeria.quochi}{francesco.rubino}{irene.russo}@ilc.cnr.it

Abstract

This paper describes a web-service system for automatic acquisition of verb subcategorization frames (SCFs) from parsed data in Italian. The system acquires SCFs in an unsupervised manner. We created two gold standards for the evaluation of the system, the first by mixing together information from two lexica (one manually created and the second automatically acquired) and manual exploration of corpus data and the other annotating data extracted from a specialized corpus (domain environment). Data filtering is accomplished by means of the maximum likelihood estimate (MLE). In addition to this, we assign to the extracted entries of the lexicon a confidence score and evaluate the extractor on domain specific data. The confidence score will allow the final user to easily select the entries of the lexicon in terms of their reliability.

Keywords: verb subcategorization frames, web services, lexicon acquisition

1. Introduction

Language Resources (LRs) are one of the key components in many NLP technologies. However, the manual creation of new LRs is costly, time consuming and prone to errors, i.e. missing information. One of the most valuable types of LRs are lexica; among these, lexica of predicate-argument structures, or sub-categorization frames (SCFs henceforth) constitute a useful tool for several tasks, such as machine translation, information retrieval etc. An SCF is the specification of the **number** and **type** of complements (both arguments and adjuncts) a word (verb, noun, adjective) can occur with. SCF lexica have found different applications in complex NLP systems (machine translation and parsing among others) as a means to add robustness. Previous works on the automatic SCFs of verbs have been conducted in English (Briscoe and Carroll, 1997; Korhonen et al., 2006), French (Messiant et al., 2008), Spanish (Alonso et al., 2007), Italian (Basili et al., 1997; Lenci et al., 2008) and German (Schulte im Walde, 2002). Different methods have been used to acquire SCF information (mainly shallow parsing and dependency parsing) with a varying number of extracted SCFs and different results in terms of precision and recall.

This paper describes a system for the automatic (unsupervised) acquisition of verbal SCFs in Italian, to be integrated in a distributed platform for the automatic creation of Language Resources. The methodology used is similar to those described in Messiant et al. (2008) and Lenci et al. (2008). The system is completely unsupervised, in the sense that it does not assume any pre-defined list of SCFs, but learns them from data instead. One of the most interesting feature of this work is the possibility the final users have to customize the results of the SCF extractor and obtaining different SCF lexica in terms of size and accuracy. The tool is made available as a web service through the PANACEA Platform[1].

2. The PANACEA Web Service for SCF Lexicon Acquisition

PANACEA is an EU-FP7 funded project with the main objective of building a platform that automates the stages involved in the acquisition and production of LRs, thus helping to cut the costs and time for their production. These in fact are still a major bottleneck for most Language Technology applications, e.g. Machine Translation. Technically, PANACEA is a platform of interoperable web services based on a set of Bioinformatics technologies developed by myGrid team within the scope of e-Science: Soaplab (Senger et al., 2003), used to deploy WSs, and Taverna (Missier et al. 2010), used for designing and running workflows. Based on the Taverna workflow manager, the platform allows users to combine different web service processors that may be distributed on different servers and can be used from various locations. An advantage of using workflows is that users do not need to install the tools nor to have deep knowledge of the technical aspects involved in all the technology they need to perform a given complex task.

Our SCF acquisition tool will thus be one of the services offered through the platform. Figure 1 below shows the SCF acquisition web service in the Taverna workflow editor. The service takes in input a (dependency parsed) text corpus (which can passed as input either directly or through a URLs), and optionally a list of verb lemmas for which SCFs will be acquired. The output SCF lexicon, encoded in XML and compliant to the Lexical Markup Framework (Francopoulo et al. 2006), is returned via a URL .

3. The IT-SCF Extractor

The IT-SCF Extractor (Extractor henceforth) takes as

[1] http://registry.elda.org/services/212

input dependency parsed data in the CoNLL format and is composed of three core modules:

a) a pattern extractor which identifies possible SCF patterns for each verb;
b) a SCF builder, which assigns a list of candidate SCFs to each verb and, finally;
c) a filter which removes SCFs that are considered incorrect.

The raw data is morpho-syntactically analyzed through the *FreeLing* suite for Italian (Padró et al.,

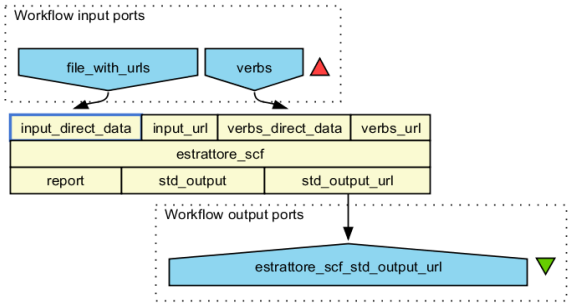

Figure 1: SCF acquisition service in the Taverna workflow editor

2010) and then parsed by the DeSR parser (Attardi and Ciaramita 2007; Attardi and Dell'Orletta 2009), through an input/output format converter (see the full workflow design in Figure 2 in appendix). The DeSR parser is one of the most accurate dependency parser for Italian (it achieved first position in both Dependency Parsing tasks at Evalita 2009). The parser builds dependency structures and chooses at each step whether to perform a shift or to create a dependency between two adjacent tokens. The dependency annotation schema is based on the ISST syntactic-functional annotation schema and does not fully distinguish between core arguments and adjuncts.

3.1 Module 1: The Pattern Extractor

The pattern extractor (PE) collects the dependencies found by the parser for each occurrence of a (target) verb. Some cases receive a special treatment, namely:

- the reflexive pronouns *si*, *mi*, *ti*, *ci* and *vi* are always extracted when they have the relations "obj" (direct object), "clit" (clitic), "comp-ind" (indirect object) and "arg" with a verb. Their presence does not give rise to a different verb entry, i.e. reflexives are not considered as separate verb entries;
- modifiers realized by adverbs, gerunds and past participles which normally are not part of an SCF of a verb are extracted and stored in a dedicated slot within the verb SCF;
- when a preposition is a dependent of the verb, the pattern extractor explores its dependent to discover the PoS which follows it (either a NP or a verbal clause in the infinitive form);
- the extraction is interrupted after a maximum of four dependent elements or when a complement clause is identified.

3.2 Module 2: The SCF Builder

The SCF builder stores the information provided by the pattern extractor as lists of eligible SCF for each verb entry. Each extracted SCF is then ordered alphabetically according to the syntactic constituents involved in order to have a normalized form of the SCF for the evaluation of the Extractor (i.e. the position of the arguments relative to to the verb is- not distincitive). Nevertheless, each occurrence of an SCF (including its frequency) is stored in a dedicated cache (SCF variants). In the lexicon, the variant with the highest frequency will be promoted as the canonical SCF form. To clarify this, let's consider the examples 1. and 2. (notice that the SCF builder output is partial, i.e. auxiliary information is not reported). The dollar symbol ($) in front of each syntactic constituent is a device to facilitate the identification of SCFs.

1. *Hanno accusato Giovanni di furto*. 'They accused Giovanni of theft'

Pattern Extractor Output: OBJ_{Giovanni}$ $COMP-DI$_{difurto}$
SCF Builder: ACCUSARE $COMP-DI_$OBJ
SCF FREQ=1 V-SCF FREQ=1
SCF Variants: ACCUSARE $OBJ_ $COMP-DI FREQ=1

2. *Hanno accusato di furto Giovanni*. 'They accused of theft Giovanni'

Pattern Extractor Output: $COMP-DIdif urto $OBJGiovanni
SCF Builder: ACCUSARE $COMP-DI_$OBJ
SCF FREQ=2 V-SCF FREQ=2
SCF Variants : ACCUSARE $COMP-DI_$OBJ FREQ=1

Due to language specific issues, i.e. the fact that Italian is a pro-drop language, and to the fact that subjects are external verb arguments, they have not been extracted at this stage of development.

3.3 Module 3: The Filter

Apart from processing errors, the output of the Extractor is noisy due to the task itself, i.e. the acquisition of verb SCFs. The most debatable issue in this task is related to the argument - adjunct distinction. Following Messiant et al. (2008), we assume that arguments tends to occur in argument position more frequently than adjuncts. Thus, frequent SCFs are assumed to be correct. The identification of these items, i.e. filtering, is accomplished in two steps by means of empirical measures based on the maximum likelihood estimate (MLE) (Korhonen et al., 2006). In this context, MLE barely corresponds to the relative frequency of the V-SCF couple. To compute

MLE we apply the formula used by Messiant et al. (2008) which is reported below:

$$MLE_{scf} = \frac{\quad}{\quad}$$

Where $|V_i - SCF_x|$ corresponds to the frequency of SCF_x with the verb Vi , i.e. the V-SCF couple, and |Vi| corresponds to the overall frequency of the verb Vi.
According to a given MLE threshold, whatever is below the empirical threshold will be rejected as probably incorrect.
In addition to this first filter, we introduce a further MLE filter, which we will call percentage on verb frequency, (PVF) for clarity's sake. Thus, for every V-SCF couple which is below the initial MLE threshold, the system reduces the length of the syntactic dependents of the SCF by taking into account all the possible combinations. Once a newly created V-SCF couple is found that already exists, then it re-assigns the associated frequency to the existing V-SCF with the highest frequency. If the updated V-SCF are above the PVF, then they are accepted, otherwise the SCF length reduction process is restarted until the V-SCF couple is above the PVF ratio.
For instance, in case we have a V-SCF couple of this kind Vx - $SCF1 $SCF2, the system splits the couple in Vx - $SCF1 and Vx -$SCF2 and assign both the frequency of the old V-SCF couple. If at least one of the newly proposed couple already exists, its assigns the frequency to the already existing frame and computes the PVF ratio. Otherwise, a new reduction process is performed until the frame is assigned.
Both the MLE and the PVF thresholds can be set by the user (they are passed as a parameter to the service), in order to allow for various types of output accuracy, depending on the specific uses the extracted data is intended for. The higher the threshold, the higher the accuracy, but obviously the lower the number of retrieved Verb-SCF pairs.
In our experiments, we established that MLE >=0.008 and PVF = 2.5% are the best filters for reaching a good balance between precision and recall (see section 4.3 below).

4. Experimental Evaluation

4.1 Data and Gold Standard

One of the most difficult issues for the evaluation of the acquisition of SCFs is related to the creation of a gold standard or, better, a lexical suite. As a matter of fact, this is an explorative task which involves also, and hopefully, the discovery of unknown SCF patterns. In order to create the gold standard, we used both existing lexical resources and manual exploration of corpus data.
Our test set was created by selecting 30 of the most frequent verbs from the Corpus *La Repubblica* (Baroni et al., 2004), which presents varied patterns in terms of semantic and syntactic features. The list of verbs is reported in Appendix A.
The first lexical resource is PAROLE/SIMPLE/CLIPS (Ruimy et al., 1998, PAROLE-IT for short). PAROLE-IT is a four-level, general purpose lexicon that has been elaborated over three different projects. The kernel of the morphological and syntactic lexicons was built in the framework of the LE-PAROLE project.
The PAROLE-IT lexicon comprises a total of 53,044 morphological units (53,044 lemmas), 37,406 syntactic units (28,111 lemmas) and 28,346 semantic units (19,216 lemmas). It was encoded in full accordance with the international standards set out in the PAROLE/SIMPLE model and based on EAGLES[2] (Leech and Wilson 1996). Its subcategorization patterns (subcat henceforth) are described in terms of optionality, syntactic function, syntagmatic realization as well as morpho-syntactic, syntactic and lexical properties of each slot filler. For the purpose of the current experiments, this information has been converted to a format which is compatible with that of the constituents used to build the SCF patterns of the system. Moreover, each optional element in the subcat has been decomposed in all its combinations. To clarify this point, let's consider example 3:

3. PAROLE SCF: *t-ppconopt-xa* [= transitive verb, with direct object and optional argument realized by a PP introduced by preposition "*con*"]

SCF Conversion: $OBJ — $OBJ $COMP-CON

All syntactic functions with the exception of the direct object are converted to their superficial form or syntagmatic realization.
In order to build a more exhaustive gold standard, another existing resource has been used: LexIt[3], which allows exploring distributional profiles of Italian nouns, verbs and adjectives automatically extracted from corpora with state-of-the-art computational linguistic methods. From this resource we extract information relative to V-SCF couples whose frequency is higher than or equal to 80. In addition to this, we exploit the associated statistical data to identify reliable and eligible SCFs. Eligible SCFs are included into the gold standard only if they were confirmed by the manual exploration.
Finally, we perform a manual exploration in order to identify missing SCFs from the two lexica and also to confirm eligible SCFs from LexIt. The manual exploration has been conducted on a context of 200 occurrences of the target verbs extracted from the *La Repubblica* Corpus. The manual exploration has been conducted by taking into account the syntagmatic

[2] http://www.ilc.cnr.it/EAGLES/
[3] http://sesia.humnet.unipi.it/lexit

realization of the verb complements and not their syntactic function. In the creation of the gold standard we applied the following rules:

- each SCF identified in the manual exploration was considered as correct;
- all SCFs from PAROLE were considered as correct;
- all SCFs extracted from LexIt were considered as correct only if they were confirmed by the manual exploration

The resulting gold standard contains 683 entries for the test verbs (and a total of 175 unique SCFs)[4].

4.2 First Evaluation

Table 1 reports the evaluation results in terms of (type) precision, (type-) recall and F-measure on the data extracted from the *La Repubblica* Corpus, at different MLE thresholds. As baseline, we consider all extracted SCFs as good (i.e. unfiltered).

Filter	#V-SCF extracted	P	R	F
unfiltered	35,526	.017	.848	.035
MLE 0.001 + PVF 3%	1,108	.390	.464	.424
MLE 0.003 + PVF 3%	647	.548	.351	.428
MLE 0.004 + PVF 3%	562	.603	.330	.426
MLE 0.005 + PVF 3%	503	.640	.308	.416
MLE 0.008 + PVF 3%	449	.679	.287	.403

Table 1: Identification of the MLE threshold and preliminary evaluation.

As the figures in Table 1 show, different filtering thresholds produce different lexica with a varied accuracy both in terms of precision and recall. In absolute terms, the system results are not satisfying due the unbalanced scores between Precision and Recall suggesting that there is still room for improvement. However, they are in line with the performances of other unsupervised methods.
A detailed error analysis showed that the following aspects contribute to the low performance of the system:

- some SCFs in the gold standard cannot be extracted due to the parser output; e.g. predicative complements realized by finite clauses or introduced by the preposition "come" are not recognized by the parser;
- some SCFs descriptors are too fine grained. For instance the firs gold standard distinguishes between $FIN-SUBJ-CHE and $FIN-CHE;
- some SCFs are rare. This suggests that in our gold standard we may have introduced instances of adjuncts in combination with real arguments;
- parsing errors bias the acquisition by introducing noise which could be filtered out.

Up to this point the SCF extractor is language independent (although not format and tagset independent), which is an interesting feature for a module of a distributed platform. In order to improve the Extractor output we need to introduced a new (language and parser-dependent) pre-filter to deal with parsing errors and to distinguish different verb types (e.g. transitive from intransitive verbs#, and then we create a different gold standard by varying the granularity of the syntactic constituents forming a SCF and their frequency with respect to the manual exploration.

4.3 Second experiment: pre-filtering and changing the gold standard composition

Our error analysis concentrated mainly on false negatives since these items are considered good SCFs by the system in the aim to develop a "best" SCF lexicon. We observe that most false negatives correspond to SCFs introduced after the manual exploration. To overcome this drawback, we conduct a further analysis of this subset of SCFs. We observed that SCF hapaxes could be collapsed into more frequent ones by deleting one or more syntactic constituents, thus signaling that they could be instances of adjuncts, and that it was possible to identify a relative frequency threshold on the basis of which manually identified SCFs could be considered as good candidates. This threshold corresponds to the 2.5% of the occurrences of the SCF with the verb. In addition to this, we collapse fine grained distinctions on mood for finite clausal complements into a single syntactic constituent, and removed all cases of SCFs which the Extractor was not able to identify since that type of information is missing from the parser output. Thus, we build a new gold standard for our test verbs which contains 443 V-SCF couples.
In Table 2 below we report the new evaluation figures (including a new evaluation with respect to the baseline):

Filter	#V-SCF extracted	P	R	F
Unfiltered	32,574	.013	.960	.026
MLE 0.008 + PVF 2.5%	485	.653	.557	.601

Table 2: Second Evaluation with pre-filter and new manual gold standard.

As the figures show, the new extractor has a better overall performance. The slight decrease in precision is balanced by an increase in recall of more than 27%. Although the new gold standard has been in part designed for the evaluation of the SCF Extractor by

[4] The goldstandard is also available in LMF XML format: http://panacea-lr.eu/system/gold_standards/SCF/PANACEA-SCF_Italian_gold_opendomain_LMF.xml

keeping into account the limits of the parser, it provides a positive feedback both on the creation procedure of the gold standard and on the granularity of the syntactic component of an SCF. However, the original gold standard set can be used as a reference gold standard for further experiments on the acquisition of SCFs in Italian in virtue of its dimensions and the results obtained.

4.3.1 Manual Exploration of False Positives

The manual exploration has been performed on 50% of the false positives (85/168). Not surprisingly, this set of data contains couples of verb-SCF with a relatively low frequency in terms of MLE ratio (ranging from 0.0072 up to 0.08). In particular, 75% (65/85) of the false positives qualifies as instances of true positives, while only 25% are incorrect ones. The error analysis has shown that the wrong SCF are due to parsing errors.

4.4 Assignment of Confidence Scores to extracted data

In order to give the final user a means to decide how to use the data (e.g. how to further filter the extracted data, or what to manually revise) confidence scores are calculated and added to the extracted lexicon entries.

As the acquisition method used does not provide intrinsic confidence values, confidence scores are calculated in an additional step relying on MLE values and on the gold standard, as described in the following paragraphs. Given an (MLE) ordered list of SCF pairs extracted from a corpus for which a gold standard is available, MLE thresholds can be found above which the precision corresponds to a given percentage. For instance we observe that taking all SCF pairs having MLE >= to 0.14, the extraction on the general domain corpus has precision = 100%, while including all SCFs having MLE >= 0.03 the precision drops to 90%. A method has thus been implemented to compute this automatically.

The algorithm goes as follows: given the extraction output ordered for decreasing MLP, the first SCF from the list is read and its precision is calculated; normally the first element will be in the gold standard, so the precision is going to be 100%. Then the second SCF is added and the new precision for the first two lines is calculated. At each run a line is added and the precision re-calculated. We are looking for precision intervals of 10%. Thus as soon as a verb-subcats N is added, which causes the precision to differ from 100% the MLE for the N-1 is recorded as the 100% threshold. The same is done for the first verb-subcats that drops the precision under >90%, and so on. In this way we can record the MLE thresholds above which precision is =100%, >90%, >80%, >70%.

These thresholds are then used as confidence values and applied by the extractor at each new extraction.

Each newly extracted verb-SCF pair is thus labeled with confidence 100% if its MLE is above the 100% threshold derived from the previously evaluated extraction, with confidence 90% if its MLE is above the 90% threshold, and so on.

4.5 Third experiment: evaluation in a Special Domain

As within the PANACEA project work and experiments are conducted on domain specific corpora, we need to test our SCF acquisition tool on the domain data in order to decide whether and how to proceed with domain adaptation. In the following we report on a first evaluation experiment and results of SCF acquisition from a domain corpus. The domain is one of the PANACEA domains, i.e. Environment. The current evaluation has been performed on 26 high frequency verbs in the domain corpus.

To this end, a domain-specific gold standard has been created by manually annotating a sample of 200 sentences per 26 high frequency verbs taken from the PANACEA Environment domain corpus for Italian. Annotation was carried out considering mainly surface structure, with no distinction between arguments and adjuncts, which is particularly critical especially in specialized domains where also verb modifiers, such as manner complements, may be relevant and highly salient. Finally, the gold standard has been compiled by applying a filter on frequency in order to exclude hapaxes and low frequency SCFs per verb[5].The final gold standard for the 26 verb lemmas contains 220 entries in total (i.e. distinct SCFs pairs).

Table 3 below reports the evaluation results on the Environment domain (ENV). Here we consider as baseline the default parameter settings based on previous experiments in the general domain.

Filter	#V-SCF extracted on ENV	P	R	F
MLE 0.008	328	0.49	0.58	0.530
MLE 0.009	297	0.53	0.56	0.546
MLE 0.01	281	0.56	0.55	0.552
MLE 0.02	191	0.71	0.45	0.555

Table 3: Domain Specific Evaluation Results

As expected, the accuracy for a specific domain is lower than in the general domain, as we expect the parser to be less accurate[6]. A thorough error analysis

[5] An LMF version will soon be generated and made available through the PANACEA website: www.panacea-lreu.

[6] It should also be noticed that the domain corpus used was automatically created through web crawling within the PANACEA

will help us adapting the extraction phase to the domain, as well as an analysis of the false positives will give us a better pulse on the actual precision; as proved with the general domain evaluation, this will likely increase considerably.

5. Conclusion and future work

In this paper we have reported on the results of a webservice system for the automatic acquisition of Italian verb subcategorization frames tested both for general domain and for a specialized domain (i.e. environment). We created two gold standards to evaluate the performance of this unsupervised system, reporting maximum likelihood estimate (MLE) for 30 verbs – 683 entries in the general domain gold standard and 26 verb - 220 entries in the specialized domain gold standard. Results from the first experiment, on a general domain corpus, show that parsing errors and the problematic issue of adjunct/ complement distinction substantially affect the acquisition quality and the reliability of the evaluation. The application of the IT-SCF-Extractor to a domain corpus shows a slight decrease in performance due to the same reasons. Also, manual exploration of false positives demonstrate how evaluation against a gold standard (or better, a reference lexicon), while useful for development purposes as it allows to improve the system with error analysis, is not optimal to assess the actual precision of the tool, and thus of the actual quality of the final outcome.
Finally, the paper describes how confidence scores are calculated and assigned to the extracted entries of the acquired SCF lexicon. The confidence score is meant to allow the final user to select in an easy way the entries of the lexicon in terms of their reliability.
In the near future we will test the system on data from another domain-specific corpus, the legal domain, and will decide what strategies to implement for a better performance on the domains, including experimenting with other statistical measures and approaches to detect domain-specific and characteristic subcategorization frames.
Experiments are ongoing for evaluating the reliability of the confidence scores assigned as described in this paper. In particular, experiments will be conducted to assess to what extent thresholds established on a given corpus and a given domain can be reliably applied to a different one. Finally, we plan to continue exploring other strategies for computing confidence scores, possibly without recourse to a gold-standard or a reference resource.

Acknowledgements

This work has been realized within the EU FP7 funded project PANACEA (Platform for Automatic. Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies) under grant agreement n. 248064.

platform, and may not necessarily be perfectly clean.

References

Alonso, L. et al. (2007). Obtaining coarse-grained classes of subcategorization patterns forSpanish. In *Proceedings of RANLP'07*.

Attardi, G. and Ciaramita M. (2007). Tree Revision Learning for Dependency Parsing, Proc. of the Human Language Technology Conference.

Attardi, G. & Dell'Orletta, F. (2009). Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: NAACL 2009*, pp. 261-264, Boulder, Colorado.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., & 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-04)*.

Basili, R., Pazienza, M.T., Vindigni, M. (1997). Corpus-driven Unsupervised Learning of Verb Subcategorization Frames. In *Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence on Advances in Artificial Intelligence*. London, UK : Springer-Verlag, pp.159-170.

Briscoe, T. & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied natural language processing*.

Francopoulo, G., et al. (2006). Lexical markup framework (LMF). In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), pages 233-236, Genoa.

Korhonen, A., Y., Krimolowsky, & T., Briscoe. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Language Resources and Evaluation (LREC'06)*.

Leech, G and Wilson, A. (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report. EAG-TCWG-MAC/R.

Lenci, A., Simonetta, M., Vito, P., & Claudia, S. (1999). Fame: a functional annotation meta-scheme for multi-modal and multi-lingual parsing evaluation. In *Proceedings of the ACL-IALL Workshop*.

Lenci, A., B., McGillivray, S., Montemagni, & V., Pirrelli (2008). Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*

Messiant, C., T., Poibeau., & A., Korhonen. (2008). Lexschem: a large subcategorization lexicon for French verbs. In *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*

Padró, L., M. Collado, S. Reese, Marina Lloberes & Irene Castellón. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*.

Ruimy, N. et al. (1998). LE PAROLE Project: the Italian Syntactic Lexicon. In *Proceedings of Euralex-98*. Liège, France.

Schulte im Walde, S. 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)*.

Appendix

In Table 4 we report the 30 verbs used for the evaluation together with their frequency in the La Repubblica Corpus.

Verb Lemma	Raw Freq.
RESTARE	124909
DIRE	409535
ANDARE	317775
ARRIVARE	121351
PARLARE	157724
CHIEDERE	124863
TORNARE	76083
VENIRE	96640
PREVEDERE	58469
PENSARE	115673
CERCARE	75485
SENTIRE	68799
PERDERE	6167
CREDERE	95092
CHIUDERE	27377
USCIRE	45140
APRIRE	40050
ENTRARE	50907
COMINCIARE	68951
TENERE	56588
CONTINUARE	102742
RENDERE	43457
EVITARE	8759
RIPETERE	25270
DIFENDERE	17302
CORRERE	28089
ACCUSARE	12780
UTILIZZARE	6472
TIRARE	19542
AMARE	28968
COMPRARE	10596

Table 4: List of verbs in the general domain gold standard

Table 5 below reports the list of verbs in the domain gold standard.

Verb Lemma
AUTORIZZARE
INTEGRARE
PORTARE
RECUPERARE
SOSTENERE
RACCOGLIERE
UTILIZZARE
CONFERIRE
PRODURRE
RENDERE
RAGGIUNGERE
OTTENERE
RIDURRE
SMALTIRE
TRATTARE
SVOLGERE
EFFETTUARE
REALIZZARE
ADOTTARE
INDIVIDUARE
CONTENERE
GARANTIRE
STABILIRE
GESTIRE
PROMUOVERE
BRUCIARE

Table 5: List of verbs in the ENV gold standard

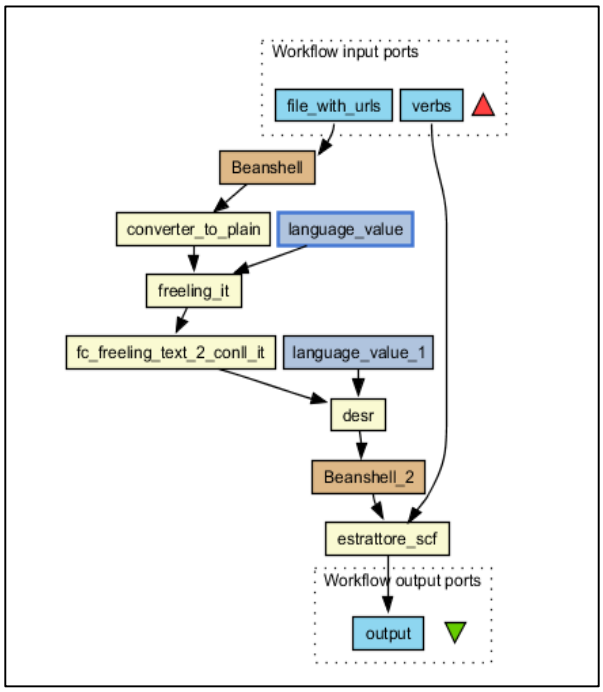

Figure 2: A workflow for SCF acquisition from