Automatic extraction of subcategorization frames for Italian

Dino Ienco, Serena Villata, Cristina Bosco

Dipartimento di Informatica - Università di Torino
Corso Svizzera 185, Torino, Italy
{ienco,villata,bosco}@di.unito.it

Abstract

Subcategorization is a kind of knowledge which can be considered as crucial in several NLP tasks, such as Information Extraction or parsing, but the collection of very large resources including subcategorization representation is difficult and time-consuming. Various experiences show that the automatic extraction can be a practical and reliable solution for acquiring such a kind of knowledge.

The aim of this paper is at investigating the relationships between subcategorization frame extraction and the nature of data from which the frames have to be extracted, e.g. how much the task can be influenced by the richness/poorness of the annotation. Therefore, we present some experiments that apply statistical subcategorization extraction methods, known in literature, on an Italian treebank that exploits a rich set of dependency relations that can be annotated at different degrees of specificity. Benefiting of the availability of relation sets that implement different granularity in the representation of relations, we evaluate our results with reference to previous works in a cross-linguistic perspective.

1. Introduction

Subcategorization specifies the number and syntactic category of verb arguments, and describes the predicate-argument structure associated with it. It is essential in various theoretical linguistic frameworks, such as Head-Driven Phrase Structure Grammar and Lexical Functional Grammar, where subcategorization consists in more or less fine-grained distinctions among verbal arguments, i.e. between complements and adjuncts, as in dependency grammars.

Moreover, verb sucategorization is a fundamental issue in several NLP tasks, for instance, in parsing where the availabity of knowledge related to Subcategorization Frames (SCFs) and the complement/adjunct distinction meaningfully increases the accuracy of results. In particular, when the language which is processed is a free word order, where complements can freely appear on the left or right side of the verbal head also mixed with adjuncts (Collins, 2003).

Therefore, the representation of subcategorization plays an important role in treebank annotation. And treebanks usually annotate subcategorization, both for free word order languages, like TIGER Corpus for German (http://www.ims.uni-stuttgart.de/projekte/TIGER/), Alpino Dependency Treebank for Dutch (http://www.let.rug.nl/˜vannoord/trees/), and Italian Syntactic Semantic Treeebank (Montemagni et al., 2003), and for fixed word order, like the English and Chinese Penn Treebanks ((Marcus et al., 1993) and (Xue, 2006)) that associate the resource with a repository, i.e. PropBank[1], where SCFs are collected.

The collection of SCFs is a very time-consuming task, in particular, because of the relative unportability of SCFs across corpora featuring different kinds of text and literary genres. Therefore, various scholars proposed the development of automatic systems for the extraction of subcategorization knowledge from linguistic corpora. For instance, meaningfull examples of works with the aim to automatically extract subcategorization frames by means of statistical methods exist for French (Chesley and Salmon-Alt, 2006), for Modern Greek (Maragoudakis et al., 2001), for Czech (Sarkar and Zeman, 2000) and for Italian (Basili et al., 1997), as we will see in Section 3.1.

In this paper, we present a set of experiments known in literature concerning the automatic extraction of SCFs from annotated sentences. On the one hand, our goal consists in investigating the complexity of the task for a free word order language, namely Italian, for which similar experiments have never previously tried [2].

On the other hand, we would like to evaluate how much and in which way the task is influenced by the features of the annotated data, from which the SCFs have to be extracted. Therefore, we selected for the development of our experiments an existing Italian treebank that features a very rich dependency-based annotation centered on a notion of predicate-argument structure, and allows for a representation of grammatical relations also scalable at different degrees of specificity.

The paper is organized as follows. In the next section, we present the data extracted from the treebank for the training. Then in the following, we describe the related works, the experiments we performed on our data and a discussion of results. We conclude with the work we planned for the next future.

2. Training data

The data set consists of 2,000 Italian sentences from a dependency-based treebank, i.e. the Turin University Treebank (TUT, download and more details at http://www.di.unito.it/˜tutreeb). The half part of the sentences included in the TUT corpus are from Civil law

[1] For PropBank see also at http://verbs.colorado.edu/˜mpalmer/projects/ace.html and http://www.cis.upenn.edu/˜chinese/

[2] Our work is distinguishable from the work of (Basili et al., 1997) by means of their use of learning techniques and clustering ones.

code, the others from newspaper articles, except for a little portion of the corpus (5%) which is from academic and novels. In the rest of this section, we describe the main features of TUT, i.e. those related to the reference language and those related to the annotation schema implemented by this resource.

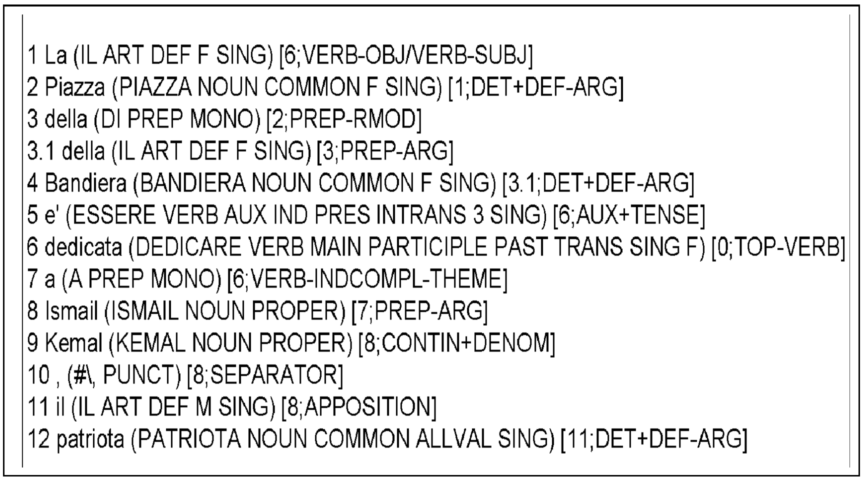
Italian is a relatively free word order language where the verb arguments do not have fixed positions. Moreover, a variety of different phrases can play the role of both complement and adjunct of the verb. For instance, in the treebank a Noun Phrase can be an adjunct like in (ALB-81) ’[Ieri mattina]$_{NP}$ sarebbe stato preso d’assalto’ - ([Yesterday morning]$_{NP}$ it would be taken by storm), or a complement like in (ALB-32) ’[Il resto del paese]$_{NP}$ era ancora sotto il controllo dell’impero’ ([The rest of the country]$_{NP}$ was still kept under control of the empire); a Prepositional Phrase can be used to introduce a complement, like in (ALB-4) ’Il Governo di Berisha appare [in difficolta]$_{PP}$’ - (The Government of Berisha seems to be [in trouble] $_{PP}$) that can also be a subordinate clause like in (ALB-89) ’Affrontando una delle piu’ gravi crisi del proprio Governo [da quando hanno sconfitto gli ex comunisti nel 1992]$_{PP}$’ - (Facing one of the most serious crises of the own Government [from when they have defeated the ex Communists in 1992]$_{PP}$), or an adjunct, like in (ALB-17) ’Tutto $\grave{e}$ cominciato [con i funerali di Artur Rustemi] $_{PP}$’ - (It is all begun [with the funerals of Artur Rustemi] $_{PP}$). As you can see below, each sentence of the TUT corpus is characterized by the indication of a subcorpus (e.g. ALB for the subcorpus on the newspaper articles from Albany) and by a progressive number (e.g. 81) that specifies the position of the sentence within the subcorpus.
In order to describe with accuracy the major features of the Italian language, TUT implements an annotation schema behind the dependency framework and following the Word Grammar theory of Hudson (Hudson, 1984). The choice of a dependency-based representation is due to the relative free word order of this language, in particular for verbal complements and adjuncts that can be distributed in the sentence in a free way that not effects the meaning of the sentence. This choice is shared by other treebanks developed for free word order languages, like the Prague Dependecny Treebank for Czech (Hajic et al., 2001), the NEGRA corpus for German (Brants et al., 1999), (Brants et al., 2002), the Alpino Treebank for Dutch (van der Beek et al., 2001), or the treebanks for Italian itself, namely the Italian Syntactic Semantic Treebank (Montemagni et al., 2003) and the Venice Italian Treebank (Delmonte et al., 2007)[3].

TUT annotation is centered upon a notion of predicate-argument structure and, therefore, systematically distinguishes and annotates various forms of complements and adjuncts (see an example in Fig. 2). Moreover, the treebank features a rich set of grammatical relations (i.e. around 250 relations) developed according to the Augmented Relational Structure (Bosco, 2004). In fact, TUT relations distinguishes and encompasses three kinds of information usually involved in grammatical relations as interrelated informational domains, called components, i.e. morpho-syntactic, functional-syntactic and semantic-syntactic. The morpho-syntactic component consists in the morphological categories of the words involved in the relation; the functional syntactic component distinguishes among a variety of dependency relations, such as SUBJ(ECT) and ARG(UMENT); the syntactic-semantic component discriminates among different kinds of adjuncts and oblique complements, such as TIME and MANNER. Valid tags for the morpho-syntactic component are 40, for the functional-syntactic are 55, and for the semantic-syntactic one they are 88 (see (Bosco, 2004) and http://www.di.unito.it/˜tutreeb).
For instance, in the example of the figures, in the relation VERB-INDCOMPL-THEME, that links the 7th word to its head, i.e. ’a’ (to), VERB is the morpho-syntactic component, INDCOMPL the functional-syntactic component, and THEME represents the syntactic-semantic one; they indicate that this is a case of an object of a verb which plays the semantic role of THEME. Another example in the same sentence is the relation DET+DEF-ARG that links ’Piazza’ (square) and ’patriota’ (patriot) with their heads; here DET+DEF is the morpho-syntactic component that specifies that the relation includes a definite determiner, while ARG is the functional-syntactic component that indicates that this is an argument od the determiner.
This allows for a representation which is scalable at different degrees of specificity. For instance, by selecting only the functional-syntactic component of each relation, we can reduce the cardinality of the relation set from 250 (fully-specified) to 74 (specified only from the functional-syntactic point of view) items.
The data selected for our experiments consist in 2,000 sentences (i.e. around 58,000 annotated tokens). They have been previously used for training of statistical methods, in particular, this same data set has been used as the development corpus for statistical parsers in a recent competition among parsing systems for Italian (Bosco et al., 2007), with results also higher to the state-of-the-art.

3. Experiments on the automatic identification of subcategorization frames

In order to overcome the problems inherent to the manual development and maintainment of large sets of SCFs, like those determined by SCFs’ variability across text genres and time, the automatic extraction of SCFs from linguistic corpora has been often applied, mainly on English, e.g., (Briscoe and Carroll, 1997) (Brent and Berwick, 1991), but also on German (Eckle and Heid, 1996), Czech (Sarkar and Zeman, 2000), Italian (Basili et al., 1997), Greek (Maragoudakis et al., 2001), and French (Chesley and Salmon-Alt, 2006).
In particular, in the perspective of our present work, it is important to analyze those results that involve Romance or free word order languages like French, Greek and Czech and based on a dependency representation of the sentences as that provided by our reference treebank.

[3] These treebanks for Italian include in a same annotation both dependencies and constituents.



```
1 La (IL ART DEF F SING) [6;VERB-OBJ/VERB-SUBJ]
2 Piazza (PIAZZA NOUN COMMON F SING) [1;DET+DEF-ARG]
3 della (DI PREP MONO) [2;PREP-RMOD]
3.1 della (IL ART DEF F SING) [3;PREP-ARG]
4 Bandiera (BANDIERA NOUN COMMON F SING) [3.1;DET+DEF-ARG]
5 e' (ESSERE VERB AUX IND PRES INTRANS 3 SING) [6;AUX+TENSE]
6 dedicata (DEDICARE VERB MAIN PARTICIPLE PAST TRANS SING F) [0;TOP-VERB]
7 a (A PREP MONO) [6;VERB-INDCOMPL-THEME]
8 Ismail (ISMAIL NOUN PROPER) [7;PREP-ARG]
9 Kemal (KEMAL NOUN PROPER) [8;CONTIN+DENOM]
10 , (#\, PUNCT) [8;SEPARATOR]
11 il (IL ART DEF M SING) [8;APPOSITION]
12 patriota (PATRIOTA NOUN COMMON ALLVAL SING) [11;DET+DEF-ARG]
```

Figure 1: TUT representation for 'La Piazza della Bandiera è dedicata a Ismail Kemal, il patriota' - *The Place of the Flag is dedicated to Ismail Kemal, the patriot.*

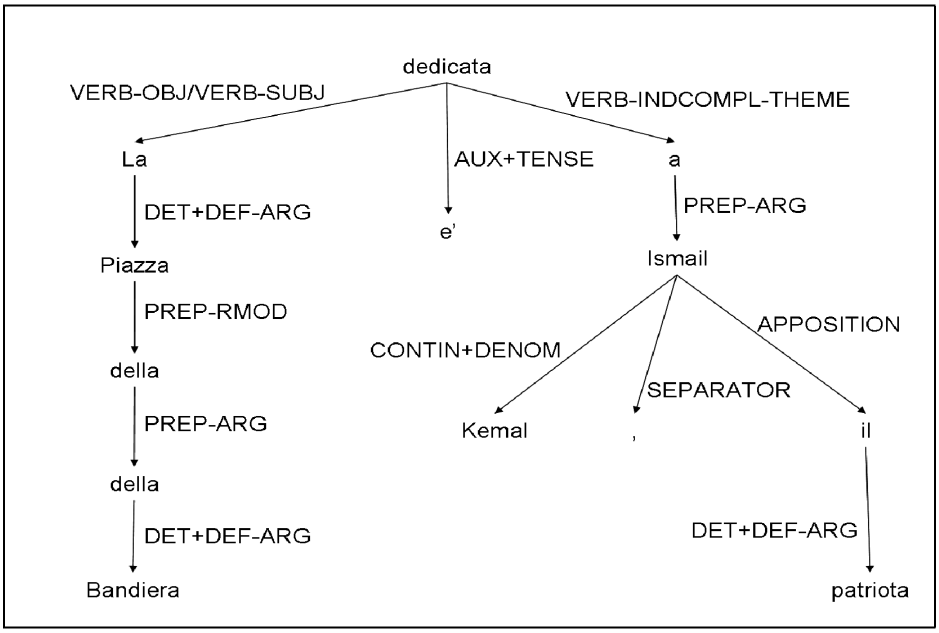


Figure 2: TUT's tree for the same sentence of Figure 1

3.1. Related works

In (Maragoudakis et al., 2001), the authors present a method for acquiring verb SCFs for Modern Greek, automatically from chunked corpora by using statistic metrics such as Log Likelihood Statistic and T-score as a measure to discover the frames. They estimate that using a free error chunker and eliminating the problem of the conjunction phrases, it is possible to achieve an accuracy higher than 75%.

In (Chesley and Salmon-Alt, 2006), the subcategorization frames for French have been acquired via VISL (a dependency-based parser), whose verb lexicon is currently incomplete with respect to subcategorization frames. The automatic extraction of French subcategorization frames is performed using binomial hypothesis testing. They obtain good results with a precision of 86.8% and a token recall rate of 54.3%. The precision is the fraction of the obtained correct SCFs divided by the total number of SCFs obtained while the recall is the fraction of the obtained correct SCFs divided by the total number of the correct SCFs.

(Sarkar and Zeman, 2000) presents some machine learning techniques for the identification of subcategorization for Czech. They compare the three different statistical techniques of T-scores, Log Likelihood Statistic and Binomial Models of Miscue Probabilities applied to this problem. The shown learning algorithm can be used to discover previously unknown SCFs from the Prague Dependency Treebank labeling dependents of a verb as either complements or adjuncts. Using these techniques, they are able to achieve 88% precision on unseen parsed text.

In all the presented cases, the automatic extraction of SCFs allows for the recovery of the real frames used in the language and provides the relative frequency of different SCFs for a given predicate.

In (Basili et al., 1997), a method for learning verb subcategorization patterns from corpora is proposed for Italian language. The basic idea of this work is that a lexicon of any quality can be used as a starting point because it is improved by corpus-driven unsupervised learning from a corpus. Conceptual clustering techniques are applied to the results of surface parsing in order to extract relevant domain typical senses and automatically build a lexicon of subcategorization frames. A core of lexico-grammatical knowledge suitable to support more sophisticated parsing strategies to be applied in a NLP application is learnt from some italian corpora.

Previous experiments for non statistics-based extraction of SCFs are also based on TUT. They are reported in (Bosco and Lombardo, 2006) where is developed a comparison among data extracted from TUT and those extracted from a manually constructed commercial Italian dictionary. In this works, 3,711 active forms are extracted from a portion of the treebank and classified in 830 lemmas. The annotation of TUT is especially centered on the predicate-argument structure and features a detailed representation of verbal complements and adjuncts. The results showed that with the relational information (i.e. dependencies) annotated in TUT 94,77% of tokens match with the data extracted from the dictionary.

3.2. Experiments

We calculated on our data measures related to the T-Score and verb environment with a variable size of the window.

3.2.1. T-Score

Making the hypothesis that the distribution of a particular frame *f* in the data is independent from the distribution of a verb *v*, we can use the T-Score statistic to detect frames highly associated to verbs. The measure of T-score can be computed by using the following equation:

$$T = \frac{p_1 - p_2}{\sqrt{\sigma^2(n_1,p_1) + \sigma^2(n_2,p_2)}}$$

It represents how much a particular frame *f* is linked with particular verb *v*. This measure can be normalized by the root of the sum of variances, as used in (Sarkar and Zeman, 2000). Based on this metrics, we observed that if the candidate *f* scored high T-Score with verb *v*, it should be considered as a valid argument of *v*.

In our experiment, we applied the measure to a version of the TUT corpus where the annotation of relations includes only the functional-syntactic component rather than the original TUT relations. We selected 50 verbs with a frequency greater than 5 occurrences and evaluate the corresponding 2,452 subcategorization frames. These 50 verbs with their 2,452 subcategorization frames are the knowledge acquired from our training set.

We tried a generalization of the knowledge acquired from the treebank on a test set composed by 226 sentences. These sentences are new with respect to the training set since not included in the TUT corpus. The half part of these sentences are from the civil law code, and the other part from italian literature for youngs. They have been selected on the basis of the fact that they include one or more occurrences of the above mentioned 50 more frequent verbs. Then they have been manually annotated according the TUT schema.

On the contrary of (Sarkar and Zeman, 2000), we do not achieve satisfactory results from this experiment. The motivations can be various, but, in our opinion mainly the following. First, we applied this evaluation on sentences from different literary genres with respect of the genres used in TUT (see section 2.). Second, our data format could be too complex because there are many relations to take in account (in the TUT schema there are 74 relationships) to be handled with trivial measures such as the T-score methods. In fact the T-Score absolute value is always under $6 * 10^{-2}$, this means that it is not statistically significant.

3.3. Experiments over Verb Environment

Using the same version of TUT corpus where only the functional-syntactic component of relations is annotated, we have also carried out a number of experiments concerning the environment of the verb. The environment of a verb can be described by windows whose dimensions represent the number and the type of dependents preceding and following the verb itself. In practice, a window *w(n, m)* represents the environment of a verb within a dependency tree, where *n* is the number of left dependents of the verb, and *m* the number of the right dependents of the verb.

We tested our data on the basis of different sizes of the verb environment: *w(-2, +3)*, i.e. two dependents preceding the verb and three dependents following it, and on *w(-3, +3)*, *w(-2, +2)*, *w(-1, +2)* and *w(-1, +1)*. For almost every environment, only a subset of these is a correct frame of the verb. We use dependents as features of the verb and the verb as a class. Over this representation, we use a Bayesian Belief Network (BBN)(Cooper and Herskovits, 1992), as used in (Kermanidis et al., 2001). A Bayesian Belief network (BBN) is a directed acyclic graphs whose nodes represent variables, and whose arcs encode conditional independencies between the variables. Nodes can represent any kind of variable, be it a measured parameter, a latent variable or a hypothesis. In fact, under conditions of uncertainty, a BBN is a relevant measure that, given a set of variables $D =< X_1, X_2...X_n >$, describes the probability distribution over this set. Each variable X_i of the set D is dependent only on its parents. The joint probability distribution over D can be computed using: $P_B(X_1...X_N) = \prod_{i=1}^{N} P(x_i|parents(X_i))$.

For this experiment, we choose 10 among the more frequent verbs that occur in TUT. In the table 4, we show the results for the window size (-2, +2) as the better representation of the environment for the selected verbs. This result positively compares with (Kermanidis et al., 2001), confirming that the window *w(-1,+2)* can be considered as an adequate dimension to capture SCFs in a relatively free word order language, like Italian or Modern Greek.

In our experiments, for instance, for a selection of verbs, we obtain the results reported in the table 5. If we analyze a particular case presented in the table in Figure 5 we can obtain the evaluation of measures of precision, recall and F-measure for each ot the verbs of this selection.

Frame	*Occurrences*
MOD MOD MOD V ARG	10
ARG MOD MOD V ARG MOD	11
MOD MOD V ARG MOD	11
ARG ARG V ARG MOD	12
ARG ARG MOD V MOD	12
ARG V ARG ARG MOD	13
MOD V	14
ARG V MOD MOD	15
ARG ARG V	16
ARG MOD V MOD	16
V MOD MOD	17
ARG V ARG MOD MOD	19
ARG MOD V	23
ARG V	24
MOD V ARG MOD	26
ARG MOD MOD V ARG	26
ARG ARG V ARG	28
ARG ARG V MOD	29
MOD MOD V ARG	29
ARG MOD V ARG ARG	33
ARG V ARG ARG	36
V ARG MOD MOD	39
MOD V ARG ARG	39
V ARG ARG MOD	43
ARG V MOD	44
ARG MOD V ARG MOD	61
MOD V ARG	86
V MOD	92
V ARG ARG	110
ARG V ARG MOD	112
empty V empty	137
ARG MOD V ARG	144
V ARG MOD	160
ARG V ARG	310
V ARG	527
ARG V MOD MOD MOD	1
MOD MOD V ARG MOD MOD	1
ARG ARG V MOD MOD	1
ARG ARG MOD V MOD MOD MOD	1
ARG ARG ARG V	1
ARG ARG ARG V MOD	1
ARG MOD MOD V ARG MOD MOD	1
ARG ARG V ARG ARG	1
MOD MOD MOD V MOD MOD	1
ARG ARG MOD V ARG MOD MOD	1
ARG MOD MOD V MOD MOD	1
MOD MOD V ARG ARG MOD	1
ARG MOD MOD V ARG ARG	2
MOD MOD V MOD	2
ARG MOD V MOD MOD MOD	2
ARG ARG V ARG MOD MOD	2
ARG ARG V MOD MOD MOD	2
MOD MOD V	2
V ARG ARG ARG	2
ARG MOD V MOD MOD	3
ARG MOD MOD V ARG ARG MOD	3
ARG ARG MOD V	5
ARG MOD MOD V MOD	5
ARG ARG MOD V MOD MOD	5
V MOD MOD MOD	5
MOD V ARG MOD MOD	5
MOD V ARG ARG MOD	5
ARG MOD V ARG ARG MOD	6
ARG ARG MOD V ARG MOD	6
ARG MOD MOD V	6
MOD MOD MOD V ARG MOD	6
MOD V MOD	7
ARG ARG V MOD MOD	8
ARG ARG MOD V ARG	9
ARG MOD V ARG MOD MOD	9
MOD MOD V ARG ARG	9

Figure 3: SCFs extracted by out system from TUT (with the only distinction between arguments ARG and modifiers MOD)

WINDOW	*BBN Acc. %*
(-1+1)	74,25 %
(-1+2)	75 %
(-2+2)	74,47 %
(-2+3)	74,01 %
(-3+3)	73,75 %

Figure 4: Application of BBN for various window size

Precision	*Recall*	*F-Measure*	*Verb*
0.96	0.923	0.941	DIVENTARE
0.684	0.433	0.531	DOVERE
0.903	0.933	0.918	APPARTENERE
0.844	0.794	0.818	RENDERE
0.805	0.917	0.857	APPLICARE
0.75	0.615	0.676	TROVARE
0.467	0.525	0.494	DIRE
0.818	0.614	0.701	TENERE
0.827	0.843	0.835	SERVIRE
0.614	0.86	0.717	STABILIRE

Figure 5: Result with environment window size (-2+2) for a selection of verbs

As an example, the table 6 shows the result of the extraction of the SCFs for a specific verb. Among the more frequent verbs of the corpus, we selected the verb stabilire (*to establish*), and we show the scores obtained by using a window *w(-3,+3).* The tabular columns give complements and adjuncts associated to the right side of the window while the tabular rows give complements and adjuncts associated to the left side. The count of each cell *[n,m]* gives the number of occurrences in which the left side (the row) appears with the right side (the column) with this verb. The total on the rows represents the number of occurrences of the selected left side with the verb stabilire and the same thing for columns. This type of computation give an idea of the distribution of complements and adjuncts for the verb taken into account, and it is similar to the work of (Ushioda et al., 1993) for English.

Finally, the table 3 shows the SCFs extracted from the corpus and their frequencies within the data set. In this extraction the relations have been further underspecified with respect to those of the native TUT data, in fact, the relations taken into account are only two, namely ARG(UMENT) and MOD(IFIER). The permutations of the elements of the left or right side are considered as the same frame because Italian, as free word order language, hasn't a fixed order for complements and modifiers in the sentences.

3.4. Discussion of results

The main difference from (Kermanidis et al., 2001), as regards our approach, is that in this case the frames are not known beforehand but are learned automatically from the training set. From the obtained result we can see that to learn models that fit correctly the distribution of the SCFs we need more information and more tagged sentences, but we can see that with the a dependency-based representation, like that implemented by TUT, we need less sentences with respect to a constituency-based representation. The results are in fact comparable to those obtained in (Maragoudakis et al., 2001) which are based on a constituency representation and by using a training set including more annotated sentences than TUT. We perform more experiments using machine learning algorithms and we obtain good accuracy with a BBN. This supply a valid result to continue our investigation in this direction.

4. Conclusions and future work

We presented a work that aims to discover the set of subcategorization frames of an Italian corpus and a number of experiments on them. These experiments can be considered as satisfactory since regardless of the small size of the corpus and the rich set of grammatical relations implemented by TUT they produced results that positively compare under some respects with the others reported in literature.

In our future work, we can extend the experiments using other traditional methods based on probability theory, for example with the one proposed in (Briscoe and Carroll, 1997).

The problem of the sparseness of the data of the treebank, which is strictly related to the small size of it, have stopped us for now from applying other statistical techniques further on the T-score measure, as seen in other approaches. A possible solution to this kind of problem can be to apply a classification of verbs of TUT, e.g. based on the verb classes of Levin (Levin, 1993). This approach can unify in a unique class verbs that occur only a few times into the corpus, thus bounding the sparseness of data.

Other directions for future work can be the investigation of the use of the knowledge provided by the SCFs in the study of text types; as attested for other languages, we can see how much the SCFs for Italian vary according to the different text categories. Moreover, we can try comparisons between the data concerning the SCFs extracted from dependency-based TUT and those extracted from its constituency-based version, like the Penn one.

5. References

R. Basili, M. T. Pazienza, and M. Vindigni. 1997. Corpus-driven unsupervised learning of verb subcategorization frames. In *Proceedings of AI*IA'97*.

C. Bosco and V. Lombardo. 2006. Comparing linguistic information in treebank annotations. In *Proceedings of LREC'06*.

C. Bosco, A. Mazzei, and V. Lombardo. 2007. Evalita parsing task: an analysis of the first parsing system contest for italian. *Intelligenza artificiale, IV - 2*.

C. Bosco. 2004. *A grammatical relation system for treebank annotation*. Ph.D. thesis, University of Turin.

T. Brants, W. Skut, and H. Uszkoreit. 1999. Syntactic annotation of a German newspaper corpus. In *ATALA Treebank Workshop*, pages 73–87.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The tiger treebank. In *Proceedings of TLT'02*.

M. R. Brent and R. C. Berwick. 1991. Automatic acquisition of sub-categorization frames from tagged text. In *Proceedings of ACL'91*.

T. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of ANLP'97*.

P. Chesley and S. Salmon-Alt. 2006. Automatic extraction of subcategorization frames for french. In *Proceedings of the LREC '06*.

M. Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4).

J. Cooper and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, (9).

R. Delmonte, A. Bristot, and S. Tonelli. 2007. VIT Venice Italian Treebank: Syntactic and quantitative features. In *Proceedings of TLT '07*.

J. Eckle and U. Heid. 1996. Extracting raw material for a german subcategorization lexicon from newspaper text. In *Proceedings of COMPLEX'96*.

J. Hajic, B. Vidova-Hladka, and P. Pajas. 2001. The Prague Dependency Treebank: Annotation structure and support. In *Proceeding of the IRCS Workshop on Linguistic Databases*.

R. Hudson. 1984. *Word Grammar*. Blackwell.

K. L. Kermanidis, M. Maragoudakis, N. Fakotakis, and G. Kokkinakis. 2001. Influence of conditional independence assumption on verb subcategorization detection. In *Proceedings of TSD'01*.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

M. Maragoudakis, K.L. Kermanidis, and G. Kokkinakis. 2001. Learning subcategorization frames from corpora: A case study for modern Greek. Technical report, Wire Communications Laboratory.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19.

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte. 2003. Building the Italian Syntactic-Semantic Treebank. In Anne Abeillé, editor, *Building and using Parsed Corpora*. Kluwer, Dordrecht.

	SUBJ/VERB	*RMOD*	*OBJ*	*RMOD OBJ*	*RMOD RMOD*	*empty*	***TOTAL***
OBJ/VERB	4	3	0	0	0	0	**7**
RMOD	1	1	0	0	0	0	**2**
SUBJ RMOD	0	0	2	0	0	0	**2**
SUBJ	0	0	0	2	2	0	**4**
empty	20	11	3	0	1	3	**38**
TOTAL	**25**	**15**	**5**	**2**	**3**	**3**	**53**

Figure 6: Occurrences of frames for verb STABILIRE (*to establish*) using w(-3,3).

A. Sarkar and D. Zeman. 2000. Automatic extraction of subcategorization frames for Czech. In *Proceedings of ACL'00*.

A. Ushioda, D. Evans, T. Gibson, and A. Waibel. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In *Proceedings of SEGLEX ACL'93*.

L. van der Beek, G. Bouma, R. Malouf, and G. van der Noord. 2001. The Alpino Dependency Treebank. In *Proceedings of CLIN'01*.

N. Xue. 2006. Annotating the predicate-argument structure of Chinese nominalizations. In *Proceedings of LREC'06*.