Using Co-Composition for Acquiring Syntactic and Semantic Subcategorisation

Pablo Gamallo **Alexandre Agustini** **Gabriel P. Lopes**

Department of Computer Science
New University of Lisbon, Portugal
{gamallo,aagustini,gpl}@di.fct.unl.pt

Abstract

Natural language parsing requires extensive lexicons containing subcategorisation information for specific sublanguages. This paper describes an unsupervised method for acquiring both syntactic and semantic subcategorisation restrictions from corpora. Special attention will be paid to the role of co-composition in the acquisition strategy. The acquired information is used for lexicon tuning and parsing improvement.

1 Introduction

Recent lexicalist Grammars project the subcategorisation information encoding in the lexicon onto syntactic structures. These grammars use accurate subcategorised lexicons to restrict potential syntactic structures. In terms of parsing development, it is broadly assumed that parsers need such information in order to reduce the number of possible analyses and, therefore, solve syntactic ambiguity. Over the last years various methods for acquiring subcategorisation information from corpora has been proposed. Some of them induce syntactic subcategorisation from tagged texts (Brent, 1993; Briscoe and Carrol, 1997; Marques, 2000). Unfortunately, syntactic information is not enough to solve structural ambiguity. Consider the following verbal phrases:

(1) [peel [$_{NP}$ the potato] [$_{PP}$ with a knife]]

(2) [peel [$_{NP}$ [$_{NP}$ the potato] [$_{PP}$ with a rough stain]]]

The attachment of "with_PP" to both the verb "peel" in phrase (1) and to the NP "the potato" in (2) does not depend only on syntactic requirements. Indeed, it is not possible to attach the PP "with a knife" to the verb "peel" by asserting that this verb subcategorises a "with_PP'. Such a subcategorisation information cannot be used to explain the analysis of phrase (2), where it is the NP "the potato" that is attached to the "with_PP". In order to decide the correct analysis in both phrases, we are helped by our world knowledge about the action of peeling, the use of knifes, and the attributes of potatoes. In general, we know that knifes are used for peeling, and potatoes can have different kinds of stains. So, the parser is able to propose a correct analysis only if the lexicon is provided with, not only syntactic subcategorisation information, but also with information on semantic-pragmatic requirements (i.e., with selection restrictions).

Other works attempt to acquire selection restrictions requiring pre-existing lexical ressources. The learning algorithm requires sample corpora to be constituted by verb-noun, noun-verb, or verb-prep-noun dependencies, where the nouns are semantically tagged by using lexical hierarchies such as WordNet (Resnik, 1997; Framis, 1995). Selection restrictions are induced by considering those dependencies associated with the same semantic tags. For instance, if verb *ratify* frequently appears with nouns semantically tagged as "legal documents" in the direct object position (e.g., *article, law, precept, ...*), then it follows that it must select for nouns denoting legal documents. Unfortunately, if a pre-defined set of semantic tags is used to annotate the training corpus, it is not obvious that the tags available are the more appropriate for extracting domain-specific semantic restrictions. If the tags were created specifically to capture corpus dependent restrictions, there could be serious problems concerning portability to a new specific domain.

By contrast, unsupervised strategies to acquire selection restrictions do not require a training corpus to be semantically annotated using pre-existing lexical hierarchies (Sekine et al., 1992; Dagan et al., 1998; Grishman and Sterling, 1994). They require only a minimum of linguistic knowledge in order to identify "meaningful" syntactic dependencies. According to the Grefenstette's terminology, they can be classified as "knowledge-poor approaches" (Grefenstette, 1994). Semantic preferences are induced by merely using co-occurrence data, i.e., by using a similarity measure to identify words which occur in the same dependencies. It is assumed that two words are semantically similar if they appear in the same contexts and syntactic dependencies. Consider for instance that the verb *ratify* frequently appear with the noun *organisation* in the subject position. Moreover, suppose that this noun turns to be similar in a particular corpus to other nouns: e.g., *secretary* and *council*. It follows that *ratify* not only selects for *organisation*, but also for its similar words. This seems to be right. However, suppose that *organisation* also appears in expressions like *the organisation of society began to be disturbed in the last decade*, or *they are involved in the actual organisation of things*, with a significant different word meaning. In this case, the noun means a particular kind of process. It seems obvious that its similar words, *secretary* and *council*, cannot appear in such subcategorisation contexts, since they are related to the other sense of the word. Soft clusters, in which words can be members of different clusters to different degrees, might solve this problem to a certain extent (Pereira et al., 1993). We claim, however, that class membership should be modeled by boolean decisions. Since subcategorisation contexts require words in boolean terms (i.e., words are either required or not required), words are either members or not members of specific subcagorisation classes. Hence, we propose a clustering method in which a word may be gathered into different boolean clusters, each cluster representing the semantic restrictions imposed by a class of subcategorisation contexts.

This paper describes an unsupervised method for acquiring information on syntactic and semantic subcategorisation from partially parsed text corpora. The main assumptions underlying our proposal will be introduced in the following section. Then, section 3 will present the different steps -extraction of candidate subcategorisation restrictions and conceptual clustering- of our learning method. In section 4, we will show how the dictionary entries are provided with the learned information. The accuracy and coverage of this information will be measured in a particular application: attachment resolution.

The experiments presented in this paper were performed on 1,5 million of words belonging to the *P.G.R.* (*Portuguese General Attorney Opinions*) corpus, which is a domain-specific Portuguese corpus containing case-law documents.

2 Underlying Assumptions

Our acquisition method is based on two theoretical assumptions. First, we assume a very general notion of linguistic subcategorisation. More precisely, we consider that in a "head-complement" dependency, not only the head imposes constraints on the complement, but also the complement imposes linguistic requirements on the head. Following Pustejovsky's terminology, we call this phenomenon "co-composition" (Pustejovsky, 1995). So, for a particular word, we attempt to learn both what kind of complements and what kind of heads it subcategorises. For instance, consider the compositional behavior of the noun *republic* in a domain-specific corpus. On the one hand, this word appears in the head position within dependencies such as *republic of Ireland, republic of Portugal*, and so on. On the other hand, it appears in the complement position in dependencies like *president of the republic, government of the republic*, etc. Given that there are interesting semantic regularities among the words cooccurring with *republic* in such linguistic contexts, we attempt to implement an algorithm letting us learn two different subcategorisation contexts:

- $[\lambda x^{\uparrow}(of; republic^{\downarrow}, x^{\uparrow})]$ where preposition of introduces a binary relation between the word *republic* in the role of "head" (role noted by arrow "$\downarrow$"), and those words that can be their "complements" (the role complement is noted by arrow "$\uparrow$"). This subcategorisation context semantically requires the complements referring to particular nations or states (indeed, only nations or states can be republics).

- $[\lambda x^{\downarrow}(of; x^{\downarrow}, republic^{\uparrow})]$ this represents a subcategorisation context that must be filled by those heads denoting specific parts of the republic: e.g., institutions, organisations, functions, and so on.

Note that the notion of subcategorisation restriction we use in this paper embraces both syntactic and semantic preferences.

The second assumption concerns the procedure for building classes of similar subcategorisation contexts. We assume, in particular, that different subcategorisation contexts are considered to be semantically similar if they have the same word distribution. Let's take, for instance, the following contexts:

$[\lambda x^{\downarrow}(of;x^{\downarrow},republic^{\uparrow})]$ $[\lambda x^{\downarrow}(of;x^{\downarrow},state^{\uparrow})]$
$[\lambda x^{\uparrow}('s;office^{\downarrow},x^{\uparrow})]$ $[\lambda x^{\uparrow}(iobj\_on;be-incumbent^{\downarrow},x^{\uparrow})]$

All of them seem to share the same semantic preferences. As these contexts require words denoting the same semantic class, they tend to possess the same word distribution. Moreover, we also assume that the set of words required by these similar subcategorisation contexts represents the extensional description of their semantic preferences. Indeed, since words *minister, president, assembly,* ... have similar distribution on those contexts, they may be used to build the extensional class of nouns that actually fill the semantic requirements of the contexts. Such words are, then, semantically subcategorised by them. Unlike most unsupervised methods to selection restrictions acquisition, we do not use the well-known strategy for measuring word similarity based on distributional hypothesis. According to this assumption, words cooccurring in similar subcategorisation contexts are semantically similar. Yet, as has been said in the Introduction, such a notion of word similarity is not sensitive to word polysemia. By contrast, the aim of our method is to measure semantic similarity between subcategorisation contexts. This allows us to assign a polysemic word to different contextual classes of subcategorisation. This strategy is also used in the Asium system (Faure and Nédellec, 1998; Faure, 2000).

3 Subcategorisation Acquisition

To evaluate the hypotheses presented above, a software package was developed to support the automatic acquisition of syntactic and semantic subcategorisation information. The learning strategy is mainly constituted by two sequential procedures. The first one aims to extract subcategorisation candidates, while the second one leads us to both identify correct subcategorisation candidates and gather them into semantic classes of subcategorisation. The two procedures will be accurately described in the remainder of the section.

3.1 Extraction of Candidates

We have developed the following procedure for extracting those syntactic patterns that could become later true subcategorisation contexts. Raw text is tagged (Marques, 2000) and then analyzed using some potentialities of the shallow parser introduced in (Rocio et al., 2001). The parser yields a single partial syntactic description of sentences, which are analyzed as sequences of basic chunks (NP, PP, VP, ...). Then, attachment is temporarily resolved by a simple heuristic based on right association (a chunk tend to attach to another chunk immediately to its right). Following our first assumption in section 2, we consider that the word heads of two attached chunks form a binary dependency that is likely to be split in two subcategorisation contexts. It can be easily seen that syntactic errors may appear since the attachment heuristic does not take into account distant dependencies.[1] For reasons of attachment errors, it is argued here that the identified subcategorisation contexts are mere hypotheses; hence they are mere subcategorisation candidates. Finally, the set of words appearing in each subcategorisation context are viewed as candidates to be a semantic class. For example, the phrase

`emanou de facto da lei`
([it] emanated in fact from the law)

[1] The errors are caused, not only due to this restrictive attachment heuristic, but also due to further misleadings, e.g., words missing from the dictionary, words incorrectly tagged, other sorts of parser limitations, etc.

would produce the following two attachments:

$$(iobj\_de; emanar^{\downarrow}, facto^{\uparrow}) \quad (de; facto^{\downarrow}, lei^{\uparrow})$$

from which the following 4 subcategorisation candidates are generated:

$$[\lambda x^{\downarrow}(iobj\_de; x^{\downarrow}, facto^{\uparrow})] \quad [\lambda x^{\uparrow}(iobj\_de; emanar^{\downarrow}, x^{\uparrow})]$$
$$[\lambda x^{\downarrow}(de; x^{\downarrow}, lei^{\uparrow})] \quad [\lambda x^{\uparrow}(de; facto^{\downarrow}, x^{\uparrow})]$$

Since the prepositional complement `de facto` represents an adverbial locution interpolated between the verb and its real complement `da lei`, the two proposed attachments are odd. Hence, the four subcategorisation contexts should not be acquired. We will see how our algorithm allows us to learn subcategorisation information that will be used later to invalidate such odd attachments and propose new ones. The algorithm basically works by comparing the similarity between the word sets associated to each subcategorisation candidate.

Let's note finally that unlike many learning approaches, information on co-composition is available for the characterization of syntactic subcategorisation contexts. In (Gamallo et al., 2001b), a strategy for measuring word similarity based on the co-composition hypothesis was compared to Grefensetette's strategy (Grefenstette, 1994). Experimental tests demonstrated that co-composition allows a finer-grained characterization of "meaningful" syntactic contexts.

3.2 Clustering Similar Contexts

According to the second assumption introduced above (section 2), two subcategorisation contexts with similar word distribution should have the same extensional definition and, then, the same selection restrictions. This way, the word sets associated with two similar contexts are merged into a more general set, which represents their extensional semantic preferences. Consider the two following subcategorisation contexts and the words that appear in them:

$$\left[\lambda x^{\uparrow}(of; infringement^{\downarrow}, x^{\uparrow})\right] = \{article\ law\ norm\ precept...\}$$
$$\left[\lambda x^{\uparrow}(dobj; infringe^{\downarrow}, x^{\uparrow})\right] = \{article\ law\ norm\ right...\}$$

Since both contexts have a similar word distribution, it can be argued that they share the same selection restrictions. Furthermore, it must be inferred that the words associated to them are all co-hyponyms belonging to the same context-dependent semantic class. In our corpus, context $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*to infringe*) is not only considered similar to context $[\lambda x^{\downarrow}(de; violação^{\downarrow}, x^{\uparrow})]$ (*infringement of*) , but also to other contexts such as: $[\lambda x^{\downarrow}(dobj; respeitar^{\downarrow}, x^{\uparrow})]$ (*to respect*) and $[\lambda x^{\uparrow}(dobj; aplicar^{\downarrow}, x^{\uparrow})]$ (*to apply*) .

In this section, we will specify the procedure for learning context-dependent semantic classes by comparing similarity between the previously extracted contextual word sets. This will be done in two steps: filtering and clustering.

3.2.1 Filtering

As has been said in the introduction, the cooperative system Asium also extract similar subcategorisation contexts (Faure and Nédellec, 1998; Faure, 2000). This system requires the interactive participation of a language specialist in order to the contextual word sets be filtered and cleaned when they are taken as input of the clustering strategy. Such a cooperative method requires manual removal of those words that have been incorrectly tagged or analyzed from the sets. Our strategy, by contrast, attempts to automatically remove incorrect words from the contextual sets. Automatic filtering requires the following subtasks:

First, each word set is associated with a list of its most similar contextual sets. Intuitively, two sets are considered as similar if they share a significant number of words. Various similarity measure coefficients were tested to create lists of similar sets. The best results were achieved using a particular weighted version of the Jaccard coefficient, where words are weighted considering both their dispersion and their relative frequency for each context (Gamallo et al., 2001a).

Then, once each contextual set has been compared to the other sets, we select the words shared by each pair of similar sets, i.e., we select the intersection between each pair of sets considered as similar. Since words that are not shared by two similar sets could be incorrect words, we remove them. Intersection allows us to clear words that are not semantically homogeneous. Thus, the intersection of two similar sets represents a class of co-hyponyms,

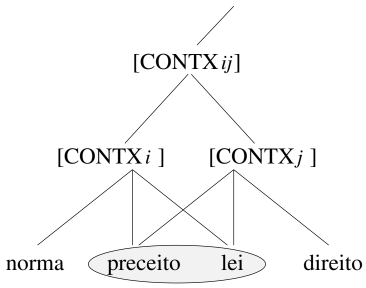

Figure 1: Clustering step

which we call *basic class*. Let's take an example. In our corpus, the most similar set extracted from $[\lambda x^{\uparrow}(de; viola\text{ção}^{\downarrow}, x^{\uparrow})]$ (*infringement of*)) is the set extracted from $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*infringe*) . Both sets share the following words:

> `sigilo princípios preceito plano norma lei estatuto disposto disposição direito`
>
> *(secret principle precept plan norm law statute disposition disposition right)*

This basic class does not contain incorrect words such as `vez, flagrantemente, obrigação, interesse` (*time, notoriously, obligation, interest*), which were oddly associated to the context $[\lambda x^{\uparrow}(de; viola\text{ção}^{\downarrow}, x^{\uparrow})]$, but which do not appear in context $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$. This class seems to be semantically homogeneous because it contains only co-hyponym words referring to legal documents. Once basic classes have been created, they are used by the conceptual clustering algorithm to build more general classes.

3.2.2 Conceptual Clustering

We use an agglomerative (bottom-up) clustering for successively aggregating the previously created basic classes. Unlike most research on conceptual clustering, aggregation does not rely on a statistical distance between classes, but on empirically set conditions and constraints (Talavera and Béjar, 1999). These conditions are discussed in (Gamallo et al., 2001a). Figure 1 shows two basic classes associated with two pairs of similar subcategorisation contexts. $[CONTX_i]$ represents a pair of similar subcategorisation contexts sharing the words `preceito, lei, norma` (*precept, law, norm*, while $[CONTX_j]$ represents another pair of similar contexts sharing the words `preceito, lei, direito` (*precept, law, right*). Both basic classes are obtained from the filtering process described in the previous section. This figure illustrates more precisely how the basic classes are aggregated into more general clusters. If two classes fill the clustering conditions, they can be merged into a new class. The two basic classes of the example are clustered into the more general class constituted by `preceito, lei, norma, direito`. At the same time, the two pairs of contexts $[CONTX_i]$ and $[CONTX_j]$ are merged into the cluster $[CONTX_{ij}]$. Such a generalization leads us to induce syntactic data that does not appear in the corpus. Indeed, we induce both that the word `norma` may appear in the syntactic contexts represented by $[CONTX_j]$, and that the word `direito` may be attached to the syntactic contexts represented by $[CONTX_i]$.

Table 1: Class Membership of `trabalho`

Cluster_1	`contrato execução exercício prazo processo procedimento trabalho` *(agreement execution practice term/time process procedure work)*
Cluster_2	`contrato exercíicio prestação recurso serviço trabalho` *(agreement practice installment appeal service work)*
Cluster_3	`actividade atribuição cargo exercício função lugar trabalho` *(activity attribution post practice function post work/job)*

3.2.3 Polysemic Words Representation

Polysemic words are placed in different clusters. For instance, consider the word `trabalho` (*work/job*). Table 1 situates this word as a member of at least three different contextual classes. Cluster_1 aggregates words referring to temporal objects. Indeed, they are co-hyponyms because they appear in subcategorisation contexts sharing the same selection restrictions: e.g., $[\lambda x^{\uparrow}(de; suspens\tilde{a}o^{\downarrow})]$, (*interruption of*), $[\lambda x^{\downarrow}(em; x^{\downarrow}, curso^{\uparrow})]$ (*in course*). Cluster_2 represents the result of an action. Such a meaning becomes salient in contexts like for instance $[\lambda x^{\uparrow}(iobj\_por; receber^{\downarrow}, x^{\uparrow})]$ (*to receive in payment for*). Indeed, the cause of receiving money is not the action of working, but the object done or the state achieved by working. Finally, Cluster_3 illustrates the more typical meaning of `trabalho`: it is a job, function or task, which can be carried out by professionals. This is why these co-

Table 2: Dictionary entries

• **abono** *(loan)*
· $[\lambda x^{\downarrow}(de; x^{\downarrow}, abono^{\uparrow})] =$
{aplicação caso fixação montante pagamento título}
(diligence case fixing amount payment bond)
· $[\lambda x^{\uparrow}(de; abono^{\downarrow}, x^{\uparrow})] =$
{ajuda despesa pensão quantia remuneração subsídio suplemento valor vencimento}
(assistance expense pension amount remuneration subsidy additional_tax value salary)
· $[\lambda x^{\downarrow}(iobj\_de; x^{\downarrow}, abono^{\uparrow})] =$
{conceder conter definir determinar fixar manter prever}
(concede comprise define determine fix maintain foresee)
• **emanar** *(emanate)*
· $[\lambda x^{\uparrow}(iobj\_de; emanar^{\downarrow}, x^{\uparrow})] =$
{alínea artigo código decreto diploma disposição estatuto legislação lei norma regulamento}
(paragraph article code decree diploma disposition statute legislation law norm regulation)
· $[\lambda x^{\uparrow}(iobj\_de; emanar^{\downarrow}, x^{\uparrow})] =$
{administração autoridade comissão conselho direcção estado governo ministro tribunal órgão}
(administration authority commission council direction state government minister tribunal organ)
• **presidente** *(president)*
· $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})] =$
{assembleia câmara comisão conselho direcção estado empresa gestão instituto região república secção tribunal }
(assembly chamber council direction state enterprise management institute region republic section tribunal)
· $[\lambda x^{\downarrow}(de; x^{\downarrow}, presidente^{\uparrow})] =$
{cargo categoria função lugar remuneração vencimento}
(post rank function place/post remuneration salary)

hyponyms can appear in subcategorisation contexts such as: $[\lambda x^{\downarrow}(de; inspector^{\downarrow})]$, (*of the inspector*), $[\lambda x^{\downarrow}(dobj; desempenhar^{\downarrow}, x^{\uparrow})]$ (*to accomplish*).

4 Application and Evaluation

The acquired classes are used in the following way. First, the lexicon is provided with subcategorisation information, and then, a second parsing cycle is performed in order to syntactic attachments be corrected.

4.1 Lexicon Update

Table 2 shows how the acquired classes are used to provide lexical entries with syntactic and semantic subcategorisation information. Each entry contains both the list of subcategorisation contexts and the list of word sets required by the syntactic contexts. As we have said before, such word sets are viewed as the extensional definition of the semantic preferences required by the subcategorisation contexts. Consider the information our system learnt for the verb `emanar` (see table 2). It syntactically subcategorises two kinds of "de-complements": the one semantically requires words referring to legal documents (`emana da lei` - *emanate from the law; law prescribes*), the other selects words referring to institutions (`emana da autoridade` - *emanate from the authority; authority proposes*). The semantic restrictions enables us to correct the odd attachments proposed by our syntactic heuristics for the phrase `emanou de facto da lei` (*emanated in fact from the law*). As word `facto` does not belong to the semantic class required by the verb in the "de-complement" position, we test the following "de-complement". As `lei` does belong, a new correct attachment is proposed.

Consider now the nouns `abono` (*loan*) and `presidente` (*president*). They subcategorise not only complements, but also different kinds of heads. For instance, the noun `abono` selects for "de-head nouns" like `fixação` (`fixação do abono` - *fixing the loan*), as well as for verbs like `fixar` in the direct object position: `fixar o abono` (*to fix the loan*).

4.2 Attachment Resolution Algorithm

The syntactic and semantic subcategorisation information provided by the lexical entries is used to check whether the subcategorisation candidates previously extracted by the parser are true attachments. The degree of efficiency in such a task may serve as a reliable evaluation for measuring the soundness of our learning strategy.

We assume the use of both a traditional chart parser (Kay, 1980) and a set of simple heuristics for identifying attachment candidates. Then, in order to improve the analysis, a "diagnosis parser" (Rocio et al., 2001) receives as input the sequences of chunks proposed as attachment candidates, checks them and raises correction procedures. Consider, for instance, the expression `editou o artigo` (*edited the article*). The diagnoser reads the sequence of chunks *VP(editar)* and *NP(artigo)*, and then proposes the attachment $(dobj; editar^{\downarrow}, artigo^{\uparrow})$ to be corrected by the system. Correction is performed by accepting or rejecting the proposed attachment. This is done looking for the subcategorisation information contained in the lexicon dictionary, information which has been acquired by the clustering method described above. Four tasks are performed to check the attachment heuristics:

Task 1a - Syntactic checking of `artigo`: check word `artigo` in the lexicon. Look for the syntactic restriction $[\lambda x^{\downarrow}(dobj; x^{\downarrow}, artigo^{\uparrow})]$. If `artigo` has this syntactic restriction, then, pass to the semantic checking. Otherwise, pass to task 2a.

Task 1b - Semantic checking of `artigo`: check the semantic restriction associated with $[\lambda x^{\downarrow}(dobj; x^{\downarrow}, artigo^{\uparrow})]$. If word `editar` belongs to that restricted class, then we can infer that $(dobj; editar^{\downarrow}, artigo^{\uparrow})$ is a binary relation. Attachment is then confirmed. Otherwise, pass to task 2a.

Task 2a - Syntactic checking of `editar`: check word `editar` in the lexicon. Look for the syntactic restriction $[\lambda x^{\uparrow}(dobj; editar^{\downarrow}, x^{\uparrow})]$. If `editar` has this syntactic restriction, then, pass to the semantic checking. Otherwise, attachment cannot be confirmed.

Task 2b - Semantic checking of `editar`: check the semantic restriction associated with $[\lambda x^{\uparrow}(dobj; editar^{\downarrow}, x^{\uparrow})]$. If word `artigo` belongs to that restricted class, then we can infer that $(dobj; editar^{\downarrow}, artigo^{\uparrow})$ is a binary relation. Attachment is then confirmed. Otherwise, attachment cannot be confirmed.

Semantic checking is based on the co-specification hypothesis stated above. According to this hypothesis, two chunks are syntactically attached only if one of these two conditions is verified: either the complement is semantically required by the head, or the head is semantically required by the complement.

4.3 Evaluating Performance of Attachment Resolution

Table 3 shows some results of the corrections proposed by the diagnosis parser. Accuracy and coverage were evaluated on three types of attachment candidates: NP-PP, VP-NP, and VP-PP. We call *accuracy* the proportion of corrections that actually correspond to true dependencies and, then, to correct attachments. *Coverage* indicates the proportion of candidate dependencies that were actually corrected. Coverage evaluation was performed by randomly selecting as test data three sets of about 100-150 occurrences of candidate attachments from the parsed corpus. Each test set only contained one type of candidate attachments. Because of low coverage, accuracy was evaluated by using larger sets of test candidates. A brief description of the evaluation results are depicted in Table 3.

Table 3: Evaluation of Attachment Resolution on NP-PP, VP-NP, and VP-PP attachment candidates

Attachment Candidate	Accuracy (%)	Coverage (%)
NP-PP	97.43	38.38
VP-NP	97.91	18.49
VP-PP	93,87	12.74
Total	96.40	23.20

Even though accuracy reaches a very promising value (about 96%), coverage merely achieves 23%. There are two main reasons for low coverage: on the one hand, the learning method needs words to have significant frequencies through the corpus; on the other hand, words are sparse through the corpus, i.e., most words of a corpus have few occurrences. However, the significant differences between the coverage for NP-PP attachments and that for verbal attachments (i.e., VP-NP and VP-PP), leads us to believe that the values reached by coverage should increase as corpus size grows. Indeed, given that verbs are less frequent than nouns, verb occurrences are still very low in a corpus containing $1,5$ millions of word occurrences. We need larger annotated corpora to improve the learning task, in particular, concerning verb subcategorisation.

5 Future Work

As we do not propose long distance attachments, our method can not be compared with other standard corpus-based approaches to attachment resolution (Hindle and Rooth, 1993; Brill and Resnik, 1994; Li and Abe, 1998). Long distance attachments only will be considered after having achieved the corrections for immediate dependencies in the first cycle of syntactic analysis. We are currently working on the specification of new analysis cycles in order to long distance attachments be solved. Consider again the phrase `emanou de facto da lei`. At the second cycle, the diagnoser proposed that the first PP `de facto` is not corrected attached to `emanou`. At the third cycle, the system will check whether the second PP `da lei` may be attached to the verb. We will perform n-cycles of attachment propositions, until no candidates are available. At the end of the process, we will be able to measure in a more accurate way what is the degree of robustness the parser may achieve.

6 Acknowledgement

This work is supported in part by grants of Fundação para a Ciência e Tecnologia, Portugal; Federal Agency for Post-Graduate Education (CAPES), Brazil; Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil; and the MLIS 4005 European project TRADAUT-PT.

References

Michael Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(3):243–262.

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *COLING*.

Ted Briscoe and John Carrol. 1997. Automatic extraction of subcategorization from corpora. In *ANCP'97*, Washington, DC, USA.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1998. Similarity-based methods of word coocurrence probabilities. *Machine Learning*, 43.

David Faure and Claire Nédellec. 1998. Asium: Learning subcategorization frames and restrictions of selection. In *ECML98, Workshop on Text Mining*.

David Faure. 2000. *Conception de méthode d'aprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph.D. thesis, Université Paris XI Orsay, Paris, France.

Francesc Ribas Framis. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin.

Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. 2001a. Selection restrictions acquisition from corpora. In *EPIA'01*, pages 30–43, Porto, Portugal. LNAI, Springer-Verlag.

Pablo Gamallo, Caroline Gasperin, Alexandre Agustini, and Gabriel P. Lopes. 2001b. Syntactic-based methods for measuring word similarity. In *TSD-2001*, pages 116–125. Berlin:Springer Verlag.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

Ralph Grishman and John Sterling. 1994. Generalizing automatically generated selectional patterns. In *COLING'94*.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Martin Kay. 1980. *Alghorith schemata and data structures in syntactic processing*. Technical report, XEROX PARK, Palo Alto, Ca., Report CSL-80-12.

Hang Li and Naoki Abe. 1998. Word clustering and disambiguation based on co-occurrence data. In *Coling-ACL'98)*, pages 749–755.

Nuno Marques. 2000. *Uma Metodologia para a Modelação Estatística da Subcategorização Verbal*. Ph.D. thesis, Univ. Nova de Lisboa, Lisboa, Portugal.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *ACL'93*, pages 183–190, Ohio.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *ACL-SIGLEX Workshop on Tagging with Lexical Semantics*, Washinton DC.

V. Rocio, E. de la Clergerie, and J.G.P. Lopes. 2001. Tabulation for multi-purpose partial parsing. *Journal of Grammars*, 4(1).

Satoshi Sekine, Jeremy Carrol, Sofia Ananiadou, and Jun'ichi Tsujii. 1992. Automatic learning for semantic collocation. In *Applied Natural Language Processing*, pages 104–110.

Luis Talavera and Javier Béjar. 1999. Integrating declarative knowledge in hierarchical clustering tasks. In *Intelligent Data Analysis*, pages 211–222.