Learning Subcategorisation Information to Model a Grammar with "Co-restrictions"

Pablo Gamallo — Alexandre Agustini — Gabriel P. Lopes

CITI
Departamento de Informática
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
2825-114 Caparica, Portugal
{gamallo,aagustini,gpl}@di.fct.unl.pt

ABSTRACT. *This paper describes two different tasks involving the notion of subcategorisation in NLP. First, it presents a specific strategy to acquire both nominal and verbal subcategorisation from text corpora. More precisely, we describe an unsupervised method for extracting syntactic and semantic subcategorisation from partially parsed texts. The second task concerns the usage of subcategorisation information for parsing. The use of this information leads to a drastic reduction of the number of arcs obtained by the parser. The notion of subcategorisation underlying these two tasks is based on a particular linguistic phenomenon, named "co-restriction". This states that two dependent words impose restrictions on each other. The main contribution of our work is to use co-restriction to define and learn subcategorisation information.*

RÉSUMÉ. *Cet article porte sur l'acquisition de la subcatégorisation et sur son usage dans des applications du TAL. Notre premier objectif consiste à élaborer une méthode permettant l'acquisition automatique, à partir d'un corpus, de la subcatégorisation nominale et verbale. En particulier, nous décrivons ici une stratégie non supervisée qui a comme objet l'extraction de la subcatégorisation syntaxique et sémantique à partir de textes partiellement analysés. Le second objectif de cet article vise à utiliser l'information apprise au préalable dans le processus d'analyse syntaxique. L'usage de cette information amène à une réduction significative du nombre d'arcs proposés par le parser. L'idée de subcatégorisation sous-jacent à notre travail est basée sur un phénomène linguistique particulier, appelé «corestriction». La corestriction consiste en ceci : chacun des mots constituant une dépendance syntaxique impose des restrictions sémantiques sur les autres mots dépendants. La contribution principale de notre travail est, précisément, l'introduction de la corestriction dans la définition de la subcatégorisation.*

KEYWORDS: *Information Extraction, Subcategorisation Acquisition, Selection Restrictions, Clustering, Co-restriction.*

MOTS-CLÉS : *extraction d'information, acquisition de la subcatégorisation, restrictions sélectionnelles, clustering, corestriction.*

1. Introduction

Lexicalised grammars need accurate subcategorised lexicons to restrict syntactic structures. In terms of parsing development, it is broadly assumed that parsers need subcategorisation information to reduce the number of possible analyses and, therefore, solve syntactic ambiguity. Over the last years various methods for acquiring subcategorisation from corpora have been proposed. Some of them induce syntactic subcategorisation from tagged texts [BRE 93, BRI 97, BUC 98, MAR 00]. Unfortunately, syntactic information is not enough to solve structural ambiguity. Structural ambiguity only can be reduced in a significant way if the parser is able to process information concerning semantico-pragmatic requirements (i.e., concerning selection restrictions). Acquisition and assessment of selection restrictions are not trivial tasks.

Several learning methods aim at acquiring selection restrictions using pre-existing lexical resources (e.g., WordNet) to annotate corpora [RES 97, FRA 95]. These external resources are considered as repositories of semantic tags. The steps of the learning procedure are the following: First, a training corpus is annotated by providing nouns (or eventually other words) with their possible semantic tags. Then, pairs $< syntactic\_position, semantic\_tag >$ are selected from the annotated corpus. Finally, the appropriateness of each pair $< syntactic\_position, semantic\_tag >$ is evaluated by means of statistical measures. For each syntactic position, the most appropriate tags represent its selection restrictions. For instance, consider that the verb *ratify* frequently appears with nouns semantically tagged as `legal_document` in the direct object position (e.g., *article, law, precept, . . .*). Statistical measures should allow to learn that the tag `legal_document` is an appropriate restriction imposed by the direct object of *ratify*. 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 most adequate for extracting domain-specific semantic restrictions. Moreover, 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 [SEK 92, DAG 98, GRI 94]. 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” [GRE 94]. Semantic preferences are induced by merely using co-occurrence data. Most of these strategies model the process of filling selection restrictions as a conditional probability. More precisely, for a particular word W and a particular syntactic position Po, the aim of the strategy is to calculate the conditional probability of the word occurring, given that position:

$$P(W|Po) \qquad [1]$$

The probabilistic information obtained by Formula 1 could be used to model syntactic combination and then attachment resolution. Such a model of attachment resolution turns to be useful for enhancing probabilistic parsers.

Nevertheless, the association between a word and a syntactic position can also be defined in non-probabilistic terms. Many approaches to formal semantics [CHA 98] usually describe this association in symbolic terms. More precisely, they model such an association as a boolean function: word W either belongs or does not belong to the requirements imposed by position Po:

$$W \in \lambda x Po(x) \qquad [2]$$

where $\lambda x Po(x)$ represents the set of words that can appear in position Po. The aim of this paper is, in fact, to model the attachment resolution process by means of the boolean association expressed by Formula 2. This formula, however, presupposes a very difficult task, namely, we must discover the *appropriate* associations between each word and each position. As this task cannot be done manually for large corpora, we could take advantage of statistical measures.

So, we do not use probabilities to model the process of filling semantico-syntactic requirements, that is, we do not represent the process of associating a word to a position in statistical terms. Probabilities and other statistical methods will be used here in a different way. They will be used to put into clusters those words that represent, in extensional terms, the semantic requirements imposed by syntactic positions. That is, statistical methods will be useful to put together those words that can fill the same syntactic positions. As a result, words will be aggregated into basic-semantic clusters with the form: $\lambda x Po(x)$.

Moreover, there is a further motivation for considering the association between a word and a position as a boolean function. As we will show in Section 4, the grammar underlying our parser merely contains symbolic information. We do not use a probabilistic parser. So, we need to learn subcategorisation information in symbolic terms.

This paper describes two different NLP tasks. First, we aim at describing an unsupervised method for acquiring information on syntactic and semantic subcategorisation from partially parsed text corpora. Second, we will define a lexicalised grammar describing syntactic and semantic subcategorisation restrictions. This grammar, together with a lexicon containing the subcategorisation restrictions learnt previously, will be used for parsing the text that was used before for learning. The article is organised as follows. In the next section (2), we will explain the basic assumptions concerning our particular notion of subcategorisation. Then, section 3 will describe the learning method. Finally, in section 4, we will show how the acquired information is used by the parser in order to disambiguate the analysis.

The experiments presented in this paper were performed on two text corpora. On the one hand, we used 1.5 million word tokens belonging to the *PGR* (*Portuguese General Attorney Opinions*) corpus. This is a domain-specific Portuguese corpus containing case-law documents. On the other hand, our learning strategy was also tested on 4.5 million words of the French part contained in the *ELIF* (*European Legislation in Force*) corpus.

2. Syntactic and Semantic Subcategorisation

Words are combined following specific linguistic constraints. The constraints imposed by a particular word in order to filter the words with which it can combine are known as *subcategorisation restrictions*. The standard notion of subcategorisation involves, at least, three different aspects: the description of morphosyntactic and semantic linguistic constraints, the distinction between *predicates* and *arguments*, and the continuum between optional and non-optional arguments. In the following subsections, we will explore and analyse these three internal aspects in order to propose a particular and original characterisation of the process of subcategorisation.

2.1. *Syntactic and Semantic Constraints*

It is broadly assumed that subcategorisation information is useful for reducing syntactic ambiguity in parsing. In some cases, it seems that very abstract morphosyntactic constraints are enough to reduce the number of possible parses. Consider the two possible analyses, (1) and (2), of the expression *cut the potato with a knife*:

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

Only analysis (1) is correct. In order the grammar to select such an analysis, we need to provide the lexicon with morphosyntactic constraints. In particular, the lexicon must assert that verb *cut* requires “with-PP” arguments. This information should prevent parser from generating analysis (2), since *with the knife* is only allowed to be attached to verb *cut*.

Unfortunately, syntactic information is not enough to solve structural ambiguity in many cases. Consider the following analysis:

(3) [cut [$_{NP}$ [$_{NP}$ the potato] [$_{PP}$ with a hole]]]

If we consider that the lexicon only associates the “with_PP” constraint to the verb *cut*, it would not be possible to allow the parser to generate analysis (3), in which it is the NP *the potato* that requires the “with_PP” argument. In order to decide the correct analysis in both phrases, we must be helped by our world knowledge concerning the action of cutting, the use of knifes, and the internal properties of potatoes. In general, we know that knifes are used for cutting, and potatoes are objects likely to have holes. 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). Selection restrictions are typically used to capture facts about the world which are generally, but not necessarily, true [AND 00].

Selection restrictions can be defined either in intensional or extensional terms. An intensional definition consists in associating a more or less complex label to a syntactic position of a given word. For instance, the label `being_a_tool` can be used to characterise those words that can appear in the "with_PP" position of verb *cut*. So, label `being_a_tool` represents the intensional requirement of *cut* in that specific syntactic position. As a correct lexical description must provide noun *knife*, but not *hole*, with label `being_a_tool`, then only the first noun satisfies the selection restrictions of *cut*. In such a case, analysis (1) becomes possible.

However, there is no agreement regarding the intensional description of selection restrictions. This description presupposes the ability to enumerate and organise the semantic labels associated to words. That is, it relies on the existence of external structured sources of lexical information (e.g., WordNet). Unfortunately, most of handcrafted lexical resources define and organise semantic classes that are not always pertinent for syntactic analysis and word combination. For instance, take the expression *the judge ratified the sentence*. In the subject position of *ratify*, it is possible to find, not only, `human_beings` like *judge*, but also `institutions` (e.g., *the government ratified the sentence*) and even `legal_documents`: e.g., *the law ratified the sentence*. It follows that the subject position of *ratify* requires words belonging to the heterogeneous semantic class defined as the reunion of: `human_beings`, `institutions`, and `legal_documents`. Unfortunately, such heterogeneous classes are not usually defined in any current lexical resource nor thesaurus.

Since the intensional definition of selection restriction is not a trivial task, we will focus on the extensional definition.

The extensional characterisation of the selection restrictions imposed by *cut* in the "with-PP" position consists in enumerating those nouns that are allowed to be there: *knife*, *buck-saw*, *hand*, etc. Note that an extensional definition does not need to provide noun *hand* with abstract labels such as `being_a_tool`. This would lead us to the well-known inconsistencies of lexical inheritance [GUA 98]: if a hand is a tool, and a tool is an artefact, then a hand is an artefact. Obviously, word *hand* should not be described as a human artefact (at least, regarding the *by default* definition). From an extensional viewpoint, we merely need to know that *hand* belongs to the set of nouns that may co-occur with *cut* in the "with-PP" position. Section 3 will deal with the process of acquiring the noun sets that represent selection restrictions in extensional terms.

2.2. *Predicate-Argument Structure*

This section introduces the properties of the predicate–argument structure of binary dependencies. We will start by introducing well-known ideas concerning this structure and the process of linguistic restriction. Then, we will make new assumptions on what we consider to be the process of restriction. These new assumptions will be used later, in Section 3, to elaborate our learning method.

2.2.1. *Simple Restriction*

It is broadly assumed that a binary syntactic dependency is constituted by both the word that imposes linguistic constraints (the predicate) and the word that must fill such constraints (its argument). In a syntactic dependency, each word plays a fixed role. The argument is perceived as the word specifying or modifying the syntactic-semantic constraints imposed by the predicate, while the latter is viewed as the word specified or modified by the former.

Besides the "predicate-argument" structure, binary dependencies also have the "head-modifier" pattern. While the former structure is related to the process of subcategorisation, the latter drives the process of transferring morphosyntactic features to higher grammatical levels.

Let's see now how these two linguistic structures ("predicate-argument" and "head-modifier") are put in relation. Typically, when the head is considered as the linguistic predicate, the modifier is conceived as the *complement* of the head. Otherwise, if the predicate turns to be the modifier, then this one is viewed as an *attribute* or an *adjunct* of the head. In other words, the modifier of a "head-modifier" structure is described as a *complement* if it satisfies the linguistic requirements of the head. By contrast, it appears to be an *attribute* or an *adjunct* if it requires the head to be appropriate to its restrictions.

The most typical case of a head viewed as a predicate is the case of a verb as head within a direct object dependency. For instance, verb *cut* can be described as a predicate that subcategorises in the direct object position a NP with some specific semantic properties. Concerning these subcategorisation restrictions, the NP *a potato* is an adequate complement.

The most typical cases of modifiers viewed as predicates are the standard uses of adjectives and adverbs. Following standard approaches to formal semantics, adjectives are unary predicates that take as arguments the denotation of nouns. So in an "adjective-noun" syntactic dependency, it is the adjective that imposes the selection restrictions on the noun. For instance, adjective *brown* denotes a predicate that requires nouns denoting colour things, and *potato* is a noun that satisfies such restrictions. Likewise, an adverb like *quickly* is considered as a predicate that selects for verbal expressions denoting actions or processes. The oddity of *?to know quickly* is motivated by the fact that the verb does not satisfy the requirements of the adverb.

Notice that lexicalised grammars only provide heads with subcategorisation information. According to these grammars, modifiers do not subcategorise. So, there is no information about the semantic subcategorisation imposed by adverbs and adjectives on their head counterparts. This situation is motivated by the fact that structural ambiguity arises only if subcategorisation information regarding the complements of verbs and nouns is not available. On the contrary, grammars do not need specific information on subcategorisation in order to correctly attach either a noun to its attributive adjective, or a verbal phrase to its adjunctive adverb. From this standpoint, semantic preferences imposed by adjectives and adverbs must be only available in the semantic interpretation of the sentence, and not during the syntactic analysis.

We will see that there is a great number of cases in which the borderline between complement and adjunct is not clear. These cases will be used to question the usually rigid characterisation of the "predicate-argument" structure. Is it only one of the two related words that imposes linguistic restrictions on the other, or do both words imposes restrictions on each other?

2.2.2. *Co-restriction*

Unlike the previous description, recent linguistic research assumes that the two words related by a syntactic dependency are mutually constrained [PUS 95, GAM 02]. Both words impose linguistic restrictions on each other. It does not exist a pre-fixed "predicate-argument" opposition. Each dependent word is at the same time both a predicate and a argument. We call such a phenomenon *co-restriction*.

Consider the expression *potato with a hole*. It seems that it is not obvious whether *hole* is either the complement or the attribute of *potato* in the "with-PP" position. If it is considered as the complement, then it is the head *potato* that should be provided with the appropriate subcategorisation information. Otherwise, it should be the attribute *hole* the word imposing subcategorisation requirements on the head. Following recent research, we claim, however, that such a radical opposition does not make sense any more. It will be assumed here that dependent expressions presuppose two subcategorisation patterns. In other words, every binary dependency is constituted by two compatible "predicate-argument" structures.

On the one hand, noun *potato* requires "with-PP" complements to be NPs denoting parts or properties of potatoes. Noun *hole* satisfies such a requirement. On the other hand, we also provide noun *hole* with subcategorisation information for the "NP-with" attribute position. Indeed, in this position, it requires head nouns denoting things that can have holes. Noun *potato* fulfils such a restriction. Note that the expression *cut with the knife* could also be considered a borderline case. Either verb *cut* requires "with-PP" complements to be NPs denoting some kind of tool, or noun *knife* selects, in the "VP-with" adjunct position, for head verbs denoting activities with agents using knifes. As we consider that the two subcategorisation patterns are compatible, we provide the lexical description of both *cut* and *knife* with the two linguistic restrictions.

Nevertheless, co-restriction is not only a pertinent way of describing borderline cases, like those of the previous examples. We consider that it should also be pertinent for typical *complement* cases (e.g., the direct object relation between verbs and nouns) as well as for typical *attribute* and *adjunct* constructions (i.e., adjective-noun and verb-adverb dependencies). Let's analyse how co-restriction works within the expression "long construction". We claim that the result of co-restriction allows the two related words to be disambiguated. Note that the two combined words convey some kind of lexical underspecification. On the one hand, *construction* seems to refer to two different entities: either the process of building or the object resulting out of the process. On the other hand, adjective *long* may refer to either spatial or temporal dimensions. Yet, the interpretation of the composite expression *long construction* is no more ambiguous. It involves the activation of a specific meaning of the two related expressions. Whereas the adjective *long* activates the processual and temporal sense of the noun, this one selects for the temporal aspect of the adjective. Both expressions, *long* and *construction* co-operate together to restrict their underspecified lexical content. Analogous examples of verb-noun dependencies are described in [GAM 02].

Therefore, according to the co-restriction assumption, two syntactically dependent expressions are no longer interpreted as a standard pair "predSo, whereas they attempt to provide verbs and nouns with information on their potential complements,icate-argument", where the predicate is the active function imposing the semantic preferences on a passive argument, which matches such preferences. On the contrary, each word of a binary dependency is perceived simultaneously as a predicate and an argument. That is, each word both imposes semantic restrictions and matches semantic requirements. When one word is interpreted as an active functor, the other is perceived as a passive argument, and conversely. Both dependent expressions are simultaneously active and passive compositional terms.

2.3. *Optional and Non-Optional Arguments*

Linguists have done a considerable effort to discriminate between subcategorised elements of words (complements) and non-subcategorised ones (adjuncts). The complement/adjunct distinction is probably one of the most unclear issues in linguistics [ALD 02]. No linguistic proposal is able to distinguish in absolute terms complements from adjuncts, e.g., is *with the knife* a complement or an adjunct of *cut*?. Most proposals may be situated in one of the two following general trends: (i) Some linguists interpolate finer distinctions between the two extremes [PUS 95]. So, between true or basic complements and completely optional adjuncts it is possible to find logic complements, pseudo-complements, and so on, which share properties of both complements and adjuncts. (ii) A most radical view is to consider the complement/adjunct opposition as a *continuum* in which it is not easy to fix borderlines between what is entirely optional and obligatory [LAN 91].

We claim that the complement/adjunct distinction remains unclear because linguists still define subcategorisation as a simple restriction process. It will be assumed here that only if subcategorisation is analysed under the co-restriction assumption, it could be possible to elucidate the internal properties of complements and adjuncts.

Figure 1 illustrates what we consider the ideal case of subcategorisation. It depicts a standard syntactic dependency between two words, the Head and the Modifier, with two "**Predicate**(Argument)" structures. On the one hand, the Head is considered a specific predicate requiring a particular type of Modifier, which is viewed here as its argument: **Head**(Modifier). On the other hand, the Modifier is also conceived

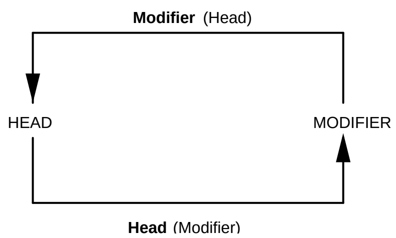

Figure 1. *Dependency without deviation*

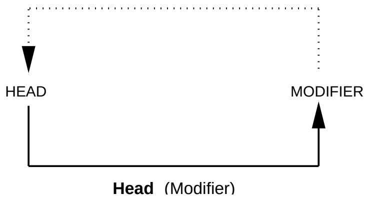

Figure 2. *Complement deviation*

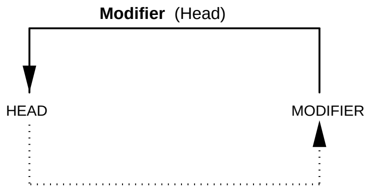

Figure 3. *Adjunct deviation*

as a predicate imposing constrains on the Head, which is here the argument: **Modifier**(Head). In the ideal situation depicted by 1, the two related words impose linguistic restrictions with the same strength. As the two subcategorisation patterns are salient, the Modifier becomes at the same time a complement and and adjunct of the Head. This situation provides a great amount of linguistic information. It supplies information on the type of word that can be the Modifier, given the specific characteristics of the Head, and conversely, we learn information on the Head, given the particular properties of the Modifier.

Figures 2 and 3 show two deviations of the ideal model depicted by 1. Figure 2 illustrates the "complement deviation" from 1. It shows the tendency of the Head of being the only predicate in the relationship. Whereas the restrictions imposed by the Head seem to be more salient, the linguistic requirements of the Modifier are put in the background. It does not mean that the Modifier is not able to require a specific type of Head. Rather, it merely means that requirements imposed by the Modifier are less significant than those imposed by the Head for this particular situation. Figure 3 illustrates the "adjunct deviation" from 1. Here, the Modifier has the tendency to be the only predicate, since its requirements appear to be more significant and, then, more salient than those of the Head.

Figure 1 represents a purely theoretical case of symmetric co-restriction. In fact, most binary dependencies imply some degree of either complement or adjunct deviation. They have the tendency to be asymmetric to a certain extent.

Nevertheless, we claim that any binary dependency should be described as a linguistic structure with two possible subcategorisation restrictions, regardless the most extreme cases of complement and adjunct deviation. We argue that co-restriction – even if it is described in asymmetric terms– must be always involved in the analysis of composite expressions. Indeed, every binary dependency presupposes any specific degree of asymmetrical co-restriction. In summary, the two predicative structures (i.e., **Head**(Modifier) and **Modifier**(Head)) underlie every binary dependency. Yet, these two structures have not often the same salience. If the **Head**(Modifier) structure turns to be the more salient one, the Modifier is conceived as a sort of complement. By contrast, If **Modifier**(Head) appears to be more salient, then the Modifier is perceived as a sort of adjunct. Both cases, complementation and adjunction, are only two obvious deviations from the ideal symmetrical situation. These two gradual deviations, however, do not prevent co-restriction to be always present and active within dependencies.

The learning strategy introduced in Section 3 takes advantage of the internal properties underlying co-restriction. We assume that any binary dependency conveys information on both **Head**(Modifier) and **Modifier**(Head) predicative patterns.

3. Acquisition Method

The method for acquiring subcategorisation information presented in this section has the following three general properties: i) it learns both syntactic and semantic subcategorisation; ii) it takes into account co-restriction; iii) it does not distinguish between optional and non–optional arguments. Before describing the different modules of the method, let us introduce first its main objectives.

3.1. *Objectives*

The main objective of our strategy is to identify the set of words required by a given word in a particular syntactic position. For instance, the verb *ratify* requires as direct object a particular set of words (e.g., *article*, *agreement*, *decree*, *law*, *statute*, etc.). This set is different from that required by the same verb as subject: e.g., *authority*, *government*, *judge*, *law*, *Portugal*, etc. The former represents those entities that can be ratified, while the latter denotes the entities that have the ability to ratify something. It is assumed here that each pair $< word_i, syntactic\_position_i >$ is associated to a set of words. This set represents, in extensional terms, the selection restrictions imposed by $word_i$ when it appears in $syntactic\_position_i$. We call the pair $< word_i, syntactic\_position_i >$ "word position". So, $< ratify, direct\_object >$ represents a particular word position of *ratify*. Later, we will define word positions in more accurate terms. Our objective will be to identify extensionally the selection restrictions subcategorised by word positions. As has been said before, selection restrictions cannot be defined in terms of ontological classes like `human_beings`, `institutions`, `vegetables`, etc. Rather, they seem to be linguistic-based classes involved in semantic composition and word combination. Consider the association:

$$< word_i, syntactic\_position_i >= word\_set_i$$

Here, there is information concerning two linguistic constraints imposed by $word_i$: first, the syntactic constraint, which is expressed by $syntactic\_position_i$; second, the semantic restriction, which is extensionally defined by $word\_set_i$. Both linguistic constraints represent the type of subcategorisation information our method tries to learn from corpora.

Another important objective of our method is to identify "similar" word positions. We assume that two word positions are considered to be semantically similar if they require the same selection restrictions. Indeed, two word positions share the same semantic restrictions if they select for similar sets of words. For instance, the direct object of *ratify* appears to require many of the words also required by the direct object of verbs like *revoke* and *approve*, or by the *of*–complement of *ratification*. So, these word positions should be defined as sharing the same selection restrictions.

Unlike most unsupervised methods for acquiring selection restrictions, we do not use the well-known strategy for measuring word similarity based on the distributional hypothesis. According to this assumption, words cooccurring in similar syntactic contexts are semantically similar. That is, according to this hypothesis, words are defined in terms of context distribution. Yet, we are not interested in learning word similarity. By contrast, the aim of our method is to measure semantic similarity between word contexts (i.e., word positions). Here, word positions are defined in terms of word distribution. The collaborative Asium system [FAU 98, FAU 00] also uses such a strategy. As will be shown later, this allows us to build clusters of word positions that share the same selection restrictions. These clusters will be extremely useful in the parsing task.

Let us also note that unlike many learning approaches, word positions are characterised taking into account the properties of co-restriction. In [GAM 01b], a standard distributional strategy for measuring word similarity using co-restriction was compared to a distributional strategy without co-restriction [GRE 94]. Experimental tests showed that co-correstriction permits a finer-grained characterization of "meaningful" syntactic positions.

Our learning strategy is mainly constituted by three sequential procedures. The aim of the first one is to extract possible word positions with their associated word sets (see subsection 3.2). The second procedure filters out the wrong words from the noisy sets extracted before (3.3). Finally, the third step consists in integrating the extracted positions and the filtered word sets into more and more general clusters. These clusters represent syntactico-semantic classes of subcategorisation (3.4). The three procedures will be accurately described in the remainder of the section.

3.2. *Extraction of Word Positions*

Concerning the Portuguese corpus, we have developed the following procedure for extracting those syntactic word positions that could be integrated later into more general clusters. First, raw text is tagged [MAR 00] and then analysed using some potentialities of the shallow parser introduced in [ROC 01]. The parser yields a single partial syntactic description of sentences, which are analysed as sequences of basic chunks (NP, PP, VP, ...). Then, attachment is temporarily solved by a simple heuristic based on right association, that is, a chunk tends to attach to another chunk immediately to its right. The two head words of two attached chunks form a binary dependency. Then, following our assumption concerning co-restriction, we split each dependency in two complementary word positions. As regards the French corpus, binary dependencies are extracted using specific attachment heuristics on the tagged text. We used the POS tagger available in [SCH 02]. Even though the texts are not partially parsed for French, the number of attachment errors is similar to that obtained by the analysis on Portuguese texts. In both cases, odd dependencies are about 25%.

Consider the French expression `livrer le tabac` (*deliver the tabac*), extracted from the *ELIF* corpus. This expression is analysed by our attachment heuristics as the following binary dependency:

$$(dobj;\, livrer^{\downarrow},\, tabac^{\uparrow}) \qquad [3]$$

Binary dependencies denote grammatical relationships between the Head and its Modifier. The word indexed by "$^{\downarrow}$" (*down* position) plays the role of Head, whereas the word indexed by "$^{\uparrow}$" (*up* position) plays the role of Modifier. In 3, *tabac* is the Modifier of the verbal Head *livrer* within the direct object (*dobj*) dependency. The syntactic dependencies we have considered are the following: subject (noted *subj*), direct object (noted *dobj*), prepositional object of verbs and nouns (both noted by the specific preposition). In this article, we will not deal with adjective and adverb adjunctions.

Following the co-restriction assumption, two word positions may be extracted from a binary dependency. We represent word positions as λ-abstractions. Moreover, each word position is associated with its "word distribution". This consists of a list of $word + frequency$ pairs, where each word must occur at least once in that position. We consider that those words appearing at least once in a word position represent, in extensional terms, the selection restrictions required by that word position. Let us see the two positions extracted from 3, and their word distributions:[1]

$$[\lambda x^{\uparrow}(dobj; livrer^{\downarrow}, x^{\uparrow})] =$$
$$\{(alcool, 1), (céréal, 1), (coton, 3), (produit, 9), (\textbf{secrétaire}, 1), (tabac, 3)...\} \; [4]$$

$$[\lambda x^{\downarrow}(dobj; x^{\downarrow}, tabac^{\uparrow})] =$$
$$\{(acheter, 5), (fabriquer, 1)(livrer, 3), (produire, 4), (vendre, 6)...\} \qquad [5]$$

In 4, it is shown the Modifier position extracted from 3, as well as those words (with their frequencies) that are the current modifiers of *livrer* in the corpus, regarding the

1. English translation of 4: *alcohol, cereal, cotton, product, secretary, tobaco*.
English translation of 5: *buy, make, delivery, produce, sell*.

direct object dependency. In 5, we show the Head position and those verbal heads (with their frequencies) cooccurring with *tabac* in the direct object relation.

Notice that we do not remove words occurring only once in a syntactic position, so as to minimize the “sparseness” problem. Linguistic corpora are sparse in the sense that most words of a corpus occur few times. So, if words with few frequencies in the corpus are not taken into account by statistical measures, the learning algorithm ends by losing pertinent linguistic information and coverage. In order to minimize coverage reduction, we will introduce non-frequent words in the sample used by the learning algorithm.

However, the fact of taking into account non-frequent words increases the amount of errors disturbing the learning task. There are several sources of noise: words missing from the dictionary, words incorrectly tagged, wrong attachments, etc. The word position illustrated by 4 occurs at least with a odd word: *secrétaire*. Note that this word has frequency 1. This enables us to keep other correct words, like *alcool* and *céréale*, which also occur only once, i.e., whose frequency is 1.

3.3. *Filtering*

The aim of the second step of our method is to remove those words that have been incorrectly associated to word positions. We call this process “filtering”. First, we calculate the “similarity” between each pair of word positions. To do it, we measure the distance between their word distributions (see below the details of this measure). Then, for each position, we select the N (where $N = 20$) most similar ones. Finally, we extract those words shared by each pair of positions considered as similar. This method is very close to the cooperative system Asium [FAU 98, FAU 00]. The difference is that our strategy is not a cooperative method requiring manual removal of those words that have been incorrectly tagged or analysed. Our strategy attempts to automatically filter out incorrect words from syntactic positions. Let’s take an example.

Consider the word position shown in 4, that is: $[\lambda x^{\uparrow}(dobj; livrer^{\downarrow}, x^{\uparrow})]$. According to our similarity measure, its word distribution is similar to that of the following positions:[2]

$$[\lambda x^{\uparrow}(de; livraison^{\downarrow}, x^{\uparrow})]\,[\lambda x^{\uparrow}(dobj; vendre^{\downarrow}, x^{\uparrow})]\,[\lambda x^{\uparrow}(de; transport^{\downarrow}, x^{\uparrow})]$$
$$[\lambda x^{\uparrow}(dobj; acheter^{\downarrow}, x^{\uparrow})]\,[\lambda x^{\uparrow}(de; achat^{\downarrow}, x^{\uparrow})]\,[\lambda x^{\uparrow}(de; importation^{\downarrow}, x^{\uparrow})]\ldots$$

Then, $[\lambda x^{\uparrow}(dobj; livrer^{\downarrow}, x^{\uparrow})]$ is put together with each one of the above positions. Finally, for each pair of similar positions, we select the words they share. For instance, take the two positions shown below in 6:

$$[\lambda x^{\uparrow}(dobj; livrer^{\downarrow}, x^{\uparrow})] \cap [\lambda x^{\uparrow}(de; livraison^{\downarrow}, x^{\uparrow})] = \{alcool, céréale, produit, tabac\} \qquad [6]$$

2. English translation: *delivery-of, to sell, transport-of, to buy, purchase-of, import-of.*

We consider that those words appearing at the same time as objects of *livrer* and *livraison* are semantically homogeneous. The set *{alcool, céréale, produit, tabac}* does not contain any more odd words such as *secrétaire*. Indeed, the fact of only extracting the words shared by two similar positions relies on two assumptions: First, we claim that those words that are not semantically required by a particular position (e.g., *secrétaire* as the direct object of *livrer*) are not likely to occur in other similar positions. Second, we also claim that the words occurring in two similar positions must be semantically required by them.

The result of associating two similar positions with their shared words forms what we call a "basic subcategorisation class". In 6 above, we join both the direct object of *livrer* and the *de*-complement of *livraison* in the same basic class. These two positions are put together into the same class of subcategorisation, since they have in common the same selection restrictions. The aim of the final learning stage will be to extend these basic classes by means of a clustering process. This will allows us to build more general subcategorisation classes.

Before explaining the final process (clustering), let us describe the measure used to calculate similarity between word positions (More details can be found in [GAM 01a]).

Various similarity measure coefficients were tested. The best results were achieved using a particular weighted version of the Lin coefficient [LIN 98]. Words are weighted considering their dispersion (global weight) and their conditional probability given a position (local weight). The weight W of a word with a position *po* is computed by equation 7:

$$W(po, word) = \log_2(P_{MLE}(word|po)) * \log_2(disp(word)) \qquad [7]$$

On the one hand, the conditional probability P_{MLE} is estimated by using the *maximum likelihood estimate (MLE)*, which is calculated in 8:

$$P_{MLE}(word|po) = \frac{f(po, word)}{F(po)} \qquad [8]$$

where $f(po, word)$ represents the frequency of $word$ appearing in *po*, and $F(po)$ is defined, for a particular word position, as the total sum of its word frequencies: $\sum_i f(po, word_i)$. On the other hand, word dispersion $disp$ is defined as the following mean:

$$disp(word) = \frac{F(word)}{number\,of\,po\,for\,word} \qquad [9]$$

The higher values of 9 are attributed to those words that are not dispersed, that is, to those words frequently appearing in few contexts. So, the Lin similarity measure sim between two positions is computed by 10[3]:

$$sim(po_1, po_2) = \frac{\sum_{common_i}(W(po_1, word_i) + W(po_2, word_i))}{\sum_j(W(po_1, word_j) + W(po_2, word_j))} \qquad [10]$$

3. $common$ means that just common words to both positions po_1 and po_2 are computed.

3.4. *Conceptual Clustering*

We use an agglomerative (bottom-up) clustering for aggregating basic subcategorisation classes into larger clusters. Unlike most research on conceptual clustering, aggregation does not rely on a statistical distance between classes, but on empirically set conditions and constraints [TAL 99]. For instance, one of these conditions states that two classes constituted by n words can be merged only if they share, at least, $n-1$ words. An additional condition states that only classes constituted by more than 4 words can be involved in the clustering process. This set of conditions, used in the first version of our clustering algorithm, is discussed more accurately in [GAM 01a]. As these conditions turned out to be too restrictive, we implemented a second version of the clustering algorithm, by using some weaker constraints: e.g., we reduced the proportion of words that two classes must share to be joined, as well as the number of words constituting them. But a new restriction was added to keep the high accuracy of the more restrictive version. A thesaurus of similar words, which has been built automatically, was used to check whether two classes can be clustered or not. These new constraints allowed us to learn more subcategorisation information.

Figure 4 shows how two basic subcategorisation classes are clustered into one more general class. For two basic classes such as:

1) the class associating the word set *{alcool, céréale, produit, tabac}* to similar positions $[\lambda x^{\uparrow}(dobj; livrer^{\downarrow}, x^{\uparrow})]$ and $[\lambda x^{\uparrow}(de; livraison^{\downarrow}, x^{\uparrow})]$ (to abbreviate, we use the shorten form: [livrer-dobj & livraison-de]),

2) the class associating the word set *{céréale, produit, tabac, sucre}* to similar positions [livrer-dobj & vendre-dobj],

we obtain the more general subcategorisation class generated by relating the word set *{alcool, céréale, produit, tabac, sucre}* to the three positions: [livrer-dobj & livraison-de & vendre-dobj]. Such a generalization leads us to induce syntactic data that does not appear in the corpus. Indeed, we induce that the word *alcool* may appear in the syntactic position [vendre-dobj]. Likewise, the word *sucre* may be attached to the position [livraison-de].

Our clustering algorithm situates a polysemic word in different clusters. The clusters a polysemic word belongs to represent its different meanings. For instance, the word `Allemagne` occurs in two different subcategorisation classes (see Table 1).[4]

Cluster_1 aggregates words referring to locations, spatial regions, or more precisely states. Indeed, these words are put together because they appear in positions requiring locations: $[\lambda x^{\uparrow}(de; nord^{\downarrow}, x^{\uparrow})]$ and $[\lambda x^{\uparrow}(en; fabriquer^{\downarrow}, x^{\uparrow})]$ (i.e., *north-of* and *made-in*).

Cluster_2 enhances a different meaning of `Allemagne`. Here, words are put together because they fill the "agentive" requirements imposed by positions such as:

4. The results obtained from both *PGR* and *ELIF* corpora can be consulted at `http://terra.di.fct.unl.pt/~agustini/restr_web`.

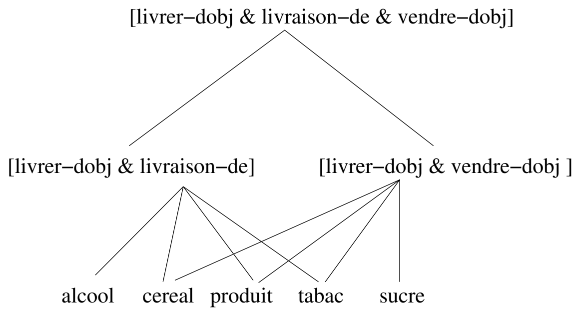

Figure 4. *Clustering step*

Cluster_1	$[\lambda x^{\uparrow}(de; nord^{\downarrow}, x^{\uparrow})] \cup [\lambda x^{\uparrow}(en; usine^{\downarrow}, x^{\uparrow})] \cup [\lambda x^{\uparrow}(en; fabriquer^{\downarrow}, x^{\uparrow})] =$ *Allemagne, Belgique, Espagne, France, Irlande, Italie, Portugal, Suède, . . .*
Cluster_2	$[\lambda x^{\uparrow}(de; discours^{\downarrow}, x^{\uparrow})] \cup [\lambda x^{\uparrow}(subj; signer^{\downarrow}, x^{\uparrow})] \cup [\lambda x^{\uparrow}(por; décider^{\downarrow}, x^{\uparrow})] =$ *Allemagne, Pays Bas, Volvo, administration, autorité, comission, conseil, . . .*

Table 1. *Two meanings of* `Allemagne`

$[\lambda x^{\uparrow}(de; discours^{\downarrow}, x^{\uparrow})]$ or $[\lambda x^{\uparrow}(subj; signer^{\downarrow}, x^{\uparrow})]$ (i.e., *discourse-of* and *to sign*). These positions have in common the fact that they require human beings and organisations being able to take decisions. According to these linguistic contexts, `Allemagne` is no more conceived as a location but as a human organisation.

3.5. *Tests and Results*

We tested our learning strategy over two corpora. On the one hand, we used a corpus with $1,643,579$ word occurrences, selected from the *P.G.R.* case-law corpus. On the other hand, we also used a larger corpus with $4,625,991$ word occurrences, which takes part of the *ELIF* text corpora. The data concerning the information extracted from these two corpora is presented in Table 2.

	Corpus P.G.R.	**Corpus ELIF**
Word Occurrences	$1,643,579$	$4,625,991$
Binary Dependencies	$267,004$	$479,631$
Word Positions	$42,032$	$71,813$
Nouns and Verbs Lemmas	$12,669$	$14,139$
Final Clusters	$2,256$	$4,653$

Table 2. *Corpus data*

Notice that the high number of clusters (see *Final Clusters* in Table 2) does not allow us to reduce them to a smaller set of general syntactico-semantic roles, such as

Agent, Patient, Theme, Instrument and so on. Yet, as they are fewer in number than the lexical entries (see *Nouns and Verbs Lemmas*), they cannot be associated with word-specific roles like, for instance, Reader, Eater, Singer, etc. The level of elaboration of these clusters can be ranged between the very abstract and the very specific lexical level. They are situated, in fact, at the domain-specific level, which is considered the more appropriate to be used in computational tasks [GIL 02].

In order to evaluate the linguistic relevance of these clusters, we will check in Section 4 if they are useful in a parsing task. The degree of efficiency in such a task (parsing) may serve as a reliable evaluation for measuring the soundness of the learning strategy.

4. Parsing

This section aims at presenting a simplified and modified version of the parser described in [ROC 01]. It is simplified because we do not deal here with the description of constituent movement. It is modified because the grammar was adapted for handling selection restrictions (that were not available then) instead of syntactic subcategorisation classes. So, in subsection 4.1 we will present the grammar that will be used next for explaining how availability of selection restrictions leads to better parses.

4.1. *Grammar*

Consider the following DCG grammar with information on "co-restriction" (see [AGU 03] for a more accurate description):

$$
\begin{aligned}
s(s(X,Y)) \;&\rightarrow\; np(X,N{+}subj{+}up{+}Cl1), && [11]\\
&\;\; vp(Y,V{+}subj{+}down{+}Cl2),\\
&\;\; \{member(N,Cl2);\\
&\;\; member(V,Cl1)\}\\
np(np(X,Y),N{+}Rel{+}Dir{+}Cl) \;&\rightarrow\; det(X), && [12]\\
&\;\; np\_nucleus(Y,N{+}Rel{+}Dir{+}Cl).\\
pp(pp(P,X),N{+}P{+}up{+}Cl) \;&\rightarrow\; prep(P), && [13]\\
&\;\; np(X,N{+}P{+}up{+}Cl).\\
vp(vp(V,X,Y,Z),V{+}subj{+}down{+}Cl) \;&\rightarrow\; v(V{+}subj{+}down{+}Cl), && [14]\\
&\;\; arg(X,V{+}Rel1{+}down{+}Cl1),\\
&\;\; arg(Y,V{+}Rel2{+}down{+}Cl2),
\end{aligned}
$$

$$np\_nucleus(np\_n(N,X,Y),N{+}Rel{+}up{+}Cl) \rightarrow \begin{array}{l} arg(Z,V{+}Rel3{+}down{+}Cl3). \\ n(N{+}Rel{+}up{+}Cl), \\ arg(X,N{+}Rel1{+}down{+}Cl1), \\ arg(Y,N{+}Rel2{+}down{+}Cl2). \end{array} \qquad [15]$$

Each grammar rule conveys at most three types of information:

– categorial organisation;

– syntactic and semantic restrictions;

– conditions on semantic adequacy;

Let us start by the information concerning the categorial organisation of linguistic groups. In rule (11), $s(X, Y)$ represents the categorial organisation of a sentence s, where X refers to the categorial organisation of the np in the subject position, while Y denotes the categorial organisation of the vp. In rule (12), $np(X, Y)$ represents the categorial organisation of a noun phrase, where X denotes the determiner and Y the categorial organisation of the noun phrase nucleus. In rule (13), $pp(P, X)$ represents the categorial organisation of a prepositional phrase, where P denotes the preposition and X the categorial organisation of the noun phrase np. In rules (14) and (15), $vp(V, X, Y, Z)$ and $np\_n(N, X, Y)$ represents the categorial organisation of vp and $np\_nucleus$, respectively. Notice that the variables X, Y, ... denote here the categorial structure of the possible arguments subcategorised by the two Head expressions: V and N. We assumed that verbs could have three arguments and nouns two. However nothing prevents us from adding more arguments.

Grammatical rules also contain information regarding the restrictions on word combination. This is the information we have extracted from corpora using the learning strategy introduced in the previous section. First, syntactico-semantic restrictions are stored in the lexicon. Then, the lexicon provides the specific constraints on word combination that will be used to instantiate the generic restrictions involved in the grammar rules. A syntactico-semantic restriction within a grammar rule has the form: $W + Rel + Dir + Cl$. Here, we associate an extensional set of words, noted Cl, with a given word: W. This word set represents extensionally the semantic restrictions W imposes in any syntactic relation Rel (e.g., in the subj(ect), dobj(ect), or whatever relation named by a specific preposition), and in any syntactic position Dir. The symbol Dir represents the "direction" of the syntactic position: it may be either the *up* direction (in case W is in the Modifier position) or the *down* direction (in case W is in the Head position). Let's see the verb restrictions contained in rule (14). $v(V + subj + down + Cl)$ enables a verb V to be associated to the class of words Cl, in the $down$ direction (i.e. in the Head position) of the $subj$(ect) relation. Furthermore, any verb, according to rules (14), (16), (17) and (18) may form a vp with 0, 1, 2, or 3 arguments. The restrictions the verb imposes on each possible argument are also introduced by rule (14). For instance, $V + Rel1 + down + Cl1$ depicts the semantic class $Cl1$ that the verb V requires, when it is the Head (i.e., in the $down$ direction) of the specific syntactic relation $Rel1$. Regarding the other possible arguments of the verb, rule (14) also contains the different word classes, $Cl2$ and $Cl3$, that V requires in the head position of other syntactic relations: $Rel2$ and $Rel3$. Given that this is a grammar with "co-restriction" information, we will find the *up* direction restrictions imposed on the verb in the argument rules (16) and (17). Within these two rules, the restrictions $N + dobj + up + Cl1$ and $N + P + up + Cl1$ introduces the possible word sets the noun requires, when it is in the Modifier position (the *up* direction) of both the $dobj$ and P (prepositional) relations. Note that rule (16) generates np arguments in the direct object position of verbs, whereas rule (17) produces pp complements of either nouns or verbs.

$$arg(X,V{+}dobj{+}down{+}Cl) \rightarrow np(X,N{+}dobj{+}up{+}Cl1), \{member(V,Cl1); lex(V{+}dobj{+}down{+}Cl), member(N,Cl))\}. \quad [16]$$

$$arg(X,W{+}P{+}down{+}Cl) \rightarrow pp(X,N{+}P{+}up{+}Cl1), \{member(W,Cl1); lex(W{+}P{+}down{+}Cl), member(N,Cl))\}. \quad [17]$$

$$arg([],\_{+}\_{+}\_{+}\_) \rightarrow []. \quad [18]$$

Finally, grammar rules also contain semantic conditions on adequacy that we need to drive and control the process of co-restriction. For instance, in rule (11) adequacy conditions are expressed by means of:

$$\{member(N, Cl2);\ member(V, Cl1)\}$$

This rule of adequacy states that the noun N and the verb V are syntactically related by the subject dependency only if either:

> N belongs to the semantic class $Cl2$, which is the class required by the verb V in the $subj$ relation, or if V belongs to the semantic class $Cl1$, which is the class required by the noun N in the $subj$ relation;

This means that two words are syntactically dependent only if at least one of them is semantically restricted by the other. So, any possible binary dependency is viewed as an adequate syntactic relation, only if, at least, either the Head is a member of the class required by the Modifier, or the Modifier is a member of the class required by the Head. In rules (16) or (17), we add the instruction $lex(W + Rel + down + Cl)$. Such an instruction enables us to initialize variable Cl by retrieving from the lex(icon) the pertinent subcategorisation information concerning the current word W. If this word does not supply adequate subcategorisation information, Cl unifies with [] (empty list: rule (18)).

According to this grammar, if there is no information about syntactic and semantic subcategorisation of nouns and verbs, then rule (18) is used and no complements are attached to them.

4.2. *Parsing with Subcategorisation*

The above grammar was compiled using DyALog system [CLE 02]. As an illustration for parsing improvement, consider the sentence:

> `O parecer do Ministro da Administração Interna do actual governo citou este projecto de regulamento em nota do seu gabinete.`
> (*The opinion of the internal affairs minister of present government cited this regulation project in a note of his cabinet.*)

If information about syntactic and semantic subcategorisation is missing, the sentence would be parsed following rule (18) as:

(4) [$_{NP}$ O parecer] [$_{PP}$ do Ministro] [$_{PP}$ da Administração Interna] [$_{PP}$ do actual governo] [$_{VP}$ citou] [$_{NP}$ este projecto] [$_{PP}$ de regulamento] [$_{PP}$ em nota] [$_{PP}$ do seu gabinete]

In this analysis there is no information on syntactic attachments. However, once the extracted subcategorisation information is available, the result is the following complete parsed sentence:

(5) [$_{S}$[$_{NP}$ O parecer [$_{PP}$ do Ministro [$_{PP}$ da Administração Interna [$_{PP}$ do actual governo]]]] [$_{VP}$ citou [$_{NP}$ este projecto [$_{PP}$ de regulamento]] [$_{PP}$ em nota [$_{PP}$ do seu gabinete]]]]]

To solve the attachments of analysis (5), the parser uses the subcategorisation restrictions shown by Table 3, which have been learnt by our acquisition method. For instance, according to these restrictions, the noun *ministro* requires, among other words, *administração* and *governo* in the "de-complement" relation. This specific information enables the parser to attach the *np* "o Ministro" to the two *pp*s: "da Administração Interna (*of internal affairs*) and "do actual governo" (*of present government*).

Table 3 also shows that the verb *citar* (*to cite*) only subcategorises a specific type of "em-complement". This allows the parser to do long distance attachment relating the *vp* "citou" (*cited*) with the *pp* "em nota" (*in a note*).

However, no information was learnt concerning the type of subject and direct object this verb subcategorises. As has been said before, the parser uses a particular type of semantic checking to solve syntactic attachment. Semantic checking is based on the co-restriction process. According to this, two phrases are syntactically attached only if one of these two conditions is verified: either the Head semantically requires the Modifier, or the Head is semantically required by the Modifier. In this particular case, it is the nominal modifier that subcategorises the verbal head. For instance, the restrictions imposed by *parecer* (*opinion*) let the vp "citar" be attached to the np "O parecer" (*The opinion*) in the subject relation. Likewise, as *projecto* (*project*) requires the verb *citar* in the $dobj$ relation, the parser is also able to attach the vp "citou" to the np "este projecto". Notice that, in these two cases, syntactic attachment is allowed by the up restrictions imposed by the verb complement (Modifier) on the own verb (Head). The use of up restrictions in parsing improvement is probably the most original contribution of our proposal.

Note also that the long distance attachment between "citou" and the pp "em nota" prevents this pp from being attached to the previous np "regulamento" (*regulation*). In fact, such an odd attachment ("regulamento em nota") was incorrectly proposed by our basic heuristics (right association), and then filtered out by the clustering strategy. In future work, we aim to apply the learning method on the new attachments proposed by the parser. This should lead to discover new subcategorisation information.

In order to explain the remaining attachments proposed by analysis (5), see again Table 3. The subcategorisation information associated to *parecer, projecto* and *nota* enables the following three np-pp attachments: "O parecer" with "do Ministro", "este projecto" with "do regulamento", and "nota" with "do seu gabinete".

Even though analysis (5) is complete, the information extracted by our learning method usually leads to partial parses. In fact, the acquired information merely allows the parsing to solve some ambiguities. Parsing will not be complete since there are other factors that may prevent such a result: i) POS tagging is not 100% error free; ii) the grammar may not be complete; and iii) for non-frequent words, probably the learning strategy is not able to learn appropriate syntactic-semantic information. Yet, even this kind of problems can also be solved by using the general methodology that is proposed by [ROC 01]. A description of this methodology, however, remains beyond the scope of this paper.

4.3. *Parsing Experiments*

We evaluated the performance of the parser presented in Section 4.1. For this purpose, we made a comparison between our method and a baseline strategy concerning the ability to propose correct syntactic attachments. As a baseline strategy, we used the attachments proposed by Right Association. For this experiment, we randomly selected 600 samples of three types of candidate attachments: $np - pp$, $vp - np$, and $vp - pp$. Since the Right Association strategy does not need any subcategorised lexicon, it always proposes a particular attachment for each one of the 600 selected

parecer (*opinion*)	$[\lambda x^{\downarrow}(de; parecer^{\downarrow}, x^{\uparrow})] =$ {administração comissão conselho entidade ministro ministério serviço órgão tribunal} (*administration commission council entity minister ministry service organ tribunal*) $[\lambda x^{\uparrow}(subj; x^{\downarrow}, parecer^{\uparrow})] =$ {citar dar prever referir solicitar} (*cite give predict refer ask_for*)
projecto (*project*)	$[\lambda x^{\downarrow}(de; projecto^{\downarrow}, x^{\uparrow})] =$ {artigo convenção código decreto despacho diploma disposição estatuto lei regulamento tratado} (*article convention code decree despatch diploma disposition statute law regulation treaty*) $[\lambda x^{\uparrow}(dobj; x^{\downarrow}, projecto^{\uparrow})] =$ {aludir citar enunciar indicar mencionar previstas qualificar referir} (*allude cite enunciate indicate mention foresee qualify refer*)
ministro (*minister*)	$[\lambda x^{\downarrow}(de; ministro^{\downarrow}, x^{\uparrow})] =$ {administração assembleia conselho entidade estado governo república órgão} (*administration assembly council entity state government republic organ*)
nota (*note*)	$[\lambda x^{\downarrow}(de; nota^{\downarrow}, x^{\uparrow})] =$ {decreto gabinete lei ministério serviço} (*decree office law ministry service*)
citar (*cite*)	$[\lambda x^{\uparrow}(em; citar^{\downarrow}, x^{\uparrow})] =$ {artigo conclusão decreto informação lei lugar norma nota número } (*article conclusion decree information law place norm note number*)

Table 3. *Constraints used to generate analysis (5)*

candidates. The general aim of this evaluation is to check whether the subcategorisation information we have learnt is adequate to be used in a parsing task.

Table 4 shows the test scores concerning precision and recall. We call *precision* the number of correct attachments suggested by the method divided by the number of total suggestions. *Recall* is computed as the number of correct attachments suggested by the method divided by the attachments that are actually correct in the sample. The recall of the baseline is 100% because the Right Association strategy is able to suggest a right attachment to all candidate dependences.

We compare the baseline method with two different versions of our system. "Conceptual Clustering Version1" contains the subcategorisation information that has been learnt by means of a very restrictive clustering algorithm, while "Conceptual Clustering Version2" is based on an algorithm provided with weaker constraints. As the clustering constraints of Version2 are are not too restrictive, the system is able to learn more lemmas with subcategorisation $(2,842)$ than the more restrictive method (Version1), which merely proposes $1,143$ new subcategorised lemmas. This implies that the total recall (23.20%) of Version1 will be lower than the recall reached by Version2: 61.34%. However, as the clustering constraints of the latter are less restrictive, the total precision of the system decreases by 4%. One of the main challenges of our current work is to tune the clustering constraints so as to reach high recall by keeping a reasonable precision. This should make the parser more efficient concerning the attachment resolution task.

BASELINE		
Attachment candidate	**Precision** (%)	**Recall** (%)
$np - pp$	83.50	100
$vp - np$	81.40	100
$vp - pp$	77.43	100
Total	80.77	100
CONCEPTUAL CLUSTERING VS1 (1,143 learned lemmas)		
Attachment candidate	**Precision** (%)	**Recall** (%)
$np - pp$	97.43	38.38
$vp - np$	97.91	18.49
$vp - pp$	93.87	12.74
Total	96.40	23.20
CONCEPTUAL CLUSTERING VS2 (2,842 learned lemmas)		
Attachment Candidate	**Precision** (%)	**Recall** (%)
$np - pp$	93.87	55.08
$vp - np$	90.90	61.72
$vp - pp$	92.72	67.54
Total	92.49	61.44

Table 4. *Evaluation of Attachment Resolution on $np - pp$, $vp - np$, and $vp - pp$ attachment candidates*

5. Conclusion and Future Work

This paper has presented a particular unsupervised strategy to automatically acquire syntactic and semantic subcategorisation restrictions. Our strategy is mainly based on two linguistic assumptions: First, it was assumed that not only the syntactic Head imposes restrictions on its Modifier, but also the latter selects for a specific type of Head. This phenomenon was called “co-restriction”. Second, we claimed that similar syntactic positions share the same selection restrictions. So, we measured, not similarity between words on the basis of their syntactic distribution, but similarity between syntactic positions on the basis of their word distribution. It was assumed that the latter kind of similarity conveys more pertinent information on linguistic subcategorisation than the former one.

The learning process allowed us to provide the lexicon with both syntactic and semantic subcategorisation information. This information was used by a parser driven by a lexicalised grammar to improve parsing accuracy. Unlike most current parsers, syntactic rules are provided with information on “co-restrictions”.

In future work, our aim is to extend the subcategorised lexicon in order to increase the coverage of the parser. For this purpose, we will use the partial results of the parser to discover new subcategorisation information. That is, the new long distance attachments identified by the parser will serve to learn more syntactic and semantic restrictions. The lexicon will be provided with these new restrictions and thereby the coverage of the parser will be increased. The successive "learning + parsing" cycles could stop as no more new information is acquired and no more new dependencies are proposed.

Acknowledgement

This work has been partially funded by Fundação para a Ciência e Tecnologia and European Commission through project MLIS 4005 TRADAUT-PT. The work by Pablo Gamallo was supported by a grant of Fundação para a Ciência e Tecnologia, Portugal (ref: PRAXIS XXI/BPD/2209/99). The work by Alexandre Agustini is supported by Federal Agency for Post-Graduate Education (CAPES), Brazil and Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil.

6. References

[AGU 03] AGUSTINI A., GAMALLO P., LOPES G. P., "Selection Restrictions Acquisition for Parsing Improvement", GESKE U., Ed., *Selected articles of 14th International Conference of Applications of Prolog (INAP2001)*, p. 129–143, Berlin:Springer Verlag, 2003.

[ALD 02] ALDEZABAL I., ARANZABE M., ATUTXA A., GOJENOLA K., SARASOLA K., "Learning Argument/Adjunt Distinction for Basque", *SIGLEX - ACL'02*, University of Pennsylvania, Philadelphia, USA, 2002, p. 42–50.

[AND 00] ANDROUTSOPOULOS I., DALE R., "Selectional Restrictions in HPSG", *18th Conference on Computational Linguistics (COLING)*, Saarbrucken, Germany, 2000, p. 15–20.

[BRE 93] BRENT M., "From grammar to lexicon: unsupervised learning of lexical syntax", *Computational Linguistics*, vol. 19, num. 3, 1993, p. 243–262.

[BRI 97] BRISCOE T., CARROL J., "Automatic Extraction of Subcategorization from Corpora", *5th Conference on Applied Natural Languague Processing (ANCP97)*, Washington, DC, USA, 1997.

[BUC 98] BUCHHOLZ S., "Unsupervised learning of subcategorisation information and its application in a parsing subtask", *10th Netherlands/Belgium Conference on Artificial Intelligence (NAIC'98)*, Amsterdam, 1998, p. 7–16.

[CHA 98] CHAMBREUIL M., ET ALII, *Sémantiques*, Cognition, Communication, Calcul, Hermès, Paris, 1998.

[CLE 02] DE LA CLERGERIE E., "Construire des analyseurs avec DyALog", *Proceedings of TALN'02*, 2002.

[DAG 98] DAGAN I., LEE L., PEREIRA F., "Similarity-Based Methods of Word Coocurrence Probabilities", *Machine Learning*, vol. 43, 1998.

[FAU 98] FAURE D., NÉDELLEC C., "ASIUM: Learning subcategorization frames and restrictions of selection", *ECML98, Workshop on Text Mining*, 1998.

[FAU 00] FAURE D., "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", PhD thesis, Université Paris XI, France, 2000.

[FRA 95] FRAMIS F. R., "On Learning More Appropriate Selectional Restrictions", *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, 1995.

[GAM 01a] GAMALLO P., AGUSTINI A., LOPES G. P., "Selection Restrictions Acquisition from Corpora", *10th Portuguese Conference on Artificial Intelligence (EPIA'01)*, Porto, Portugal, 2001, LNAI, Springer-Verlag, p. 30–43.

[GAM 01b] GAMALLO P., GASPERIN C., AGUSTINI A., LOPES G. P., "Syntactic-Based Methods for Measuring Word Similarity", MAUTNER V., MOUCEK R., MOUCEK K., Eds., *Text, Speech, and Discourse (TSD-2001)*, p. 116–125, Berlin:Springer Verlag, 2001.

[GAM 02] GAMALLO P., AGUSTINI A., LOPES G. P., "Using Co-composition for Acquiring Syntactic and Semantic Subcategorisation", *ACL-SIGLEX'02*, Philadelphia, USA, 2002.

[GIL 02] GILDEA D., JURAFSKY D., "Automatic Labeling of Semantic Roles", *Computational Linguistics*, vol. 28, num. 3, 2002, p. 245–288.

[GRE 94] GREFENSTETTE G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, USA, 1994.

[GRI 94] GRISHMAN R., STERLING J., "Generalizing Automatically Generated Selectional Patterns", *Proceedings of the 15th International on Computational Linguistics (COLING-94)*, 1994.

[GUA 98] GUARINO N., "Some Ontological Principles for Designing Upper Level Lexical Resources", *First Conference on Language Resources and Evaluations*, Spain, 1998.

[LAN 91] LANGACKER R. W., *Foundations of Cognitive Grammar: Descriptive Applications*, vol. 2, Stanford University Press, Stanford, 1991.

[LIN 98] LIN D., "Automatic Retrieval and Clustering of Similar Words", *COLING-ACL'98*, Montreal, 1998.

[MAR 00] MARQUES N., "Uma Metodologia para a Modelação Estatística da Subcategorização Verbal", PhD thesis, Universidade Nova de Lisboa, Lisboa, Portugal, 2000.

[PUS 95] PUSTEJOVSKY J., *The Generative Lexicon*, MIT Press, Cambridge, 1995.

[RES 97] RESNIK P., "Selectional Preference and Sense Disambiguation", *ACL-SIGLEX Workshop on Tagging with Lexical Semantics*, Washinton DC, 1997.

[ROC 01] ROCIO V., DE LA CLERGERIE E., LOPES J., "Tabulation for multi-purpose partial parsing", *Journal of Grammars*, vol. 4, num. 1, 2001.

[SCH 02] SCHIMD H., "TreeTagger", *A language independent part-of-speech tagger*, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html, 2002.

[SEK 92] SEKINE S., CARROL J., ANANIADOU S., TSUJII J., "Automatic Learning for Semantic Collocation", *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 1992, p. 104-110.

[TAL 99] TALAVERA L., BÉJAR J., "Integrating Declarative Knowledge in Hierarchical Clustering Tasks", *Intelligent Data Analysis*, 1999, p. 211–222.