Using Syntactic-Based Methods for Extracting Semantic Information

Pablo Gamallo[1] *, Caroline Gasperin[2], Alexandre Agustini[1] **, and Gabriel P. Lopes[1]

[1] CENTRIA, Departamento de Informática
Universidade Nova de Lisboa, Portugal
[2] Faculdade de Informática - PPGCC - PUCRS, Brasil

Abstract. This paperchapter proposes a sound characterisation of the syntactic features used to acquire semantic information from annotated text corpora. This characterisation is based on the co-specification hypothesis. According to this hypothesis, two syntactically related words impose semantic restrictions to each other. In order to study the behaviour of co-specification in different learning tasks, this paperchapter describes how the syntactic features defined in basis of co-specification could be used to appropriately learn both lists of similar words, and classes of selection restrictions

1 Introduction

The general aim of this paper is to describe the role of syntactic features in the automatic extraction of semantic information from corpora. We assume here that semantic extraction strategies need accurate and well-defined syntactic features so as to appropriately acquire semantic information.

The strategies for extracting semantic information from corpora can be roughly divided into two categories, knowledge-rich and knowledge-poor methods, according to the amount of knowledge they presuppose [5,4]. Knowledge-rich approaches require some sort of previously encoded semantic information [?,?,3,8]: domain-dependent knowledge structures, semantic tagged training copora, and/or semantic resources such as handcrafted thesauri: Roget's thesaurus, WordNet, and so on. Therefore, knowledge-rich approaches may inherit the main shortcomings and limitations of man-made lexical resources: limited vocabulary size, since they can include unnecessary general words, or do not include necessary domain-specific ones; unclear classification criteria, since their word classification is sometimes too coarse and does not provide sufficient distinction between words, or is sometimes unnecessarily detailed; and, obviously, considerable time and effort required by building thesauri by hand. By contrast, knowledge-poor approaches use no presupposed semantic knowledge for automatically extracting semantic information. These techniques can be characterised as follows: no domain- specific information is available, no semantic tagging is used, and no static sources as dictionaries or thesauri are required. They use the frequency of co-occurrences of words within various linguistic contexts (either syntactic constructions or sequences of $n-grams$) in order to extract semantic information such as word similarity [?,5,6], and selection restrictions [?,?,?].

Since these methods do not require previously defined semantic knowledge, they overcome the well-known drawbacks associated with handcrafted thesauri and supervised strategies.

According to the nature of linguistic contexts, two specific knowledge-poor strategies can also be distinguished: windows-based and syntactic-based techniques. Windows-based techniques consider an arbitrary number of words around a given word as forming its window, i.e., its context. The linguistic information about part-of-speech categories and syntactic groupings is not taken into account to characterise word contexts [1,11]. The syntactic-based strategy, on the contrary, requires specific linguistic information to specify the word context. First, it requires a part-of-speech tagger for assigning a morphosyntactic label to each word of the corpus. Then, the tagged corpus is segmented into a sequence of basic phrasal groupings (or chunks). Finally, simple attachment heuristics are used to specify the relations between and within the phrasal groupings. Once the syntactic analysis of the corpus is reached, each word in the corpus is associated to a set of syntactic contexts. Semantic information is extracted by identifying some regularities in the syntactic distribution of different words [5,6,2].

Both strategies, window-based and syntactic-based techniques, lie on the Harris' *distributional hypothesis*. According to this assumption, words occurring in similar contexts are considered as semantically similar. Usually, the similarity measure between two words is obtained by using their conditional distributions in all

* This work was supported by the PRAXIS XXI project, Fundação para a Ciência e a Tecnologia, Ministério da Ciência e a Tecnologia, Portugal.
** Research sponsored by CAPES and PUCRS - Brasil.

contexts. Even though knowledge-poor strategies may difer in the statistical definition of both *conditional distribution* and *similarity measure*, the paper will not focus on the comparative analysis of these statistical notions for semantic information extraction.

We assume that syntactic analysis opens up a much wider range of more precise distributional contexts than does simple windows strategy. As syntactic contexts represent linguistic dependencies involving specific semantic relationships, they should be considered as fine-grained clues for identifying semantically related words.

Since syntactic contexts can be defined in different ways, syntactic-based approaches can also be significantly different. Different pieces of linguistic information can be taken into account to characterise syntactic contexts. For instance, the information used by Lin [6] to define the notion of syntactic context is not the same than that used by Grefenstette [5]. Nevertheless, the choice of a particular type of syntactic context for extracting semantic information is not usually properly justified.

This way, the main objective of this paper is to define a specific notion of syntactic context. The appropriateness or the inadequacy of this definition will be tested in two different semantic extraction approaches: word similarity extraction and selection restrictions acquistion. The paper is organised as follows. In section **??**, syntactic contexts will be described on the basis of the notion of linguistic co-specification. This definition will be compared to other notions of syntactic contexts. Special attention will be paid to the syntactic contexts used by Grefenstette. Then, in section **??**, we will test if our notion of syntactic context is more appropriate than other notions for a particular task, namely, word similarity extraction. For this purpose, we will compare the results obtained using our notion of context to the results achived by using the Grefenstette's contexts [5]. Finally, in section **??**, the syntactic contexts we have defined will be used for a different task, namely the acquisition of the selection restrictions that words impose on the words with which they cooccur. It will be claimed that the accurate information contained in these syntactic contexts is not independent of the rules governing word combination. Our approach will be compared to some elements of the Faure's strategy ([**?**].

The two learning strategies for acquiring both word similarity and selection restrictions will be tested over the domain-specific text corpora *P.G.R.*[1] The fact of using specialised text corpora makes easier the learning task, given that we have to deal with a limited vocabulary with reduced polysemy. Furthermore, since these strategies are not dependent of a specific language such a Portuguese, they could be applied to different natural languages.

2 Co-specification and Syntactic Contexts

We argue that it is not possible to correctly model the acquisition of linguistic information from corpora if we are not sensitive to the co-specification hypothesis. We will define first the notion of co-specification, and then, this notion will be used to characterise and extract syntactic contexts. At the end of this section, we will compare the syntactic contexts based on co-specification to the contexts defined on the basis of simple specification.

2.1 Co-specification between Predicate-Argument Pairs

Traditionally, a binary syntactic relationship 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 relationship, each word plays a fixed role. The argument is perceived as the word specifying or modifying the syntactic-semantic constraints imposed by predicate, while the latter is viewed as the word specified or modified by the former. However, recent linguistic research assumes that the two words related by a syntactic dependency are mutually determined. Each word imposes semantic conditions on the other word of the dependency, and each word elaborates them. Consider the relationship between the polysemic verb *load* and the polysemic noun books in the not ambiguous expression *to load the books*. On the one hand, the polysemic verb *load* conveys at least two alternate meanings: "bringing something to a location" (e.g., *Ann loaded the hay onto the truck*), and "modifying a location with something" (e.g., *Ann loaded the truck with the hay*). This verb is disambiguated by taking into account the sense of the words with which it combines within the sentence. On the other hand, the noun *book(s)* is also a polysemic expression. Indeed, it refers to different types of entities: "physical objects" (*rectangular book*), and "symbolic entities" (*naive book*). Yet, the constraints imposed by the words with which it combines allows the noun to be disambiguated. Whereas the adjective *rectangular* activates the physical sense of *book*, the adjective *naive* makes reference to its symbolic content.

[1] P.G.R. (*Portuguese General Ashtorney Opinions*) is constituted by case-law documents in Portuguese (`http:coluna.di.fct.unl.py~pgr`).

In *to load the books*, the verb *load* activates the physical sense of the noun, while *books* leads *load* to refer to the event of bringing something to a location. The interpretation of the complex expression is no more ambiguous. Both expressions, *load* and *books*, cooperate to mutually restrict their meaning. The process of mutual restriction between two related words is called by Pustejovsky "co-specification" or "co-composition" [?,?]. Co-specification is based on the following idea. Two syntactically dependent expressions are no longer interpreted as a standard pair "predicate-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. Unlike most work on selection restrictions learning, our notion of "predicate-argument" frame relies on the active process of semantic co-specification, and not on the trivial operation of argument specification. This trivial operation only permits the one-way specification and disambiguation of the argument by taking into account the sense of the predicate. Any specification and disambiguation of the predicate by the argument is not considered.

In the following subsection, we will define the notion of syntactic context on the basis of the notion of co-specification.

2.2 Identification of Binary Dependencies and Extraction of Syntactic Contexts

According to the co-specification hypothesis, two dependent words can be analysed as two syntactic contexts of specification. In this subsection, we start by defining the internal structure of a dependence relationship between two words (or "binary dependence"), and then, we describe how syntactic contexts are extracted from binary dependencies.

2.3 Binary Dependencies

We assume that basic syntactic contexts are extracted from binary syntactic dependencies. Let's describe the internal structure of a dependency between two words. A syntactic dependency consists of two words and the hypothetical grammatical relationship between them. We represent a dependency as the following binary predication:

$$(r; w1^{\downarrow}, w2^{\uparrow})$$

This binary predication is constituted by the following entities:

- the binary predicate r, wich can be associated to specific prepositions, subject relations, direct object relations, etc.;
- the roles of the predicate, "$\downarrow$" and "$\uparrow$", which represent the *head* and *complement* roles, respectively;
- the two words holding the binary relation: $w1$ and $w2$.

Binary dependencies denote grammatical relationships between the head and its complement. The word indexed by "$\downarrow$" plays the role of *head*, whereas the word indexed by "$\uparrow$" plays the role of *complement*. Therefore, $w1$ is perceived as the head and $w2$ as the complement.

The binary dependencies (i.e., grammatical relationships) we have considered are the following: subject (noted *subj*), direct object (noted *dobj*), adjective modifiers (noted *modif*, prepositional object of verbs, and prepositional object of nouns, both noted by the specific preposition. Consider Table 1. The left column contains expressions constituted by two words syntactically related. The right column contains the binary dependencies used to represent these related expressions.

2.4 Extraction of Syntactic Contexts and Co-Specification

Syntactic contexts are abstract configurations of specific binary dependencies. We use λ-abstraction to represent the extraction of syntactic contexts. A syntactic context is extracted by λ-abstracting one of the related words of a binary dependency. Thus, two complementary syntactic context can be λ-abstracted from the binary predication associated with a syntactic dependency:

$$[\lambda x^{\downarrow}(r; x^{\downarrow}, w2^{\uparrow})] \qquad [\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$$

Table 1. Binary dependencies identified from related expressions

Related Expressions	Binary Dependencies
`presidente da república` (*president of the republic*)	$(de; presidente^{\downarrow}, república^{\uparrow})$
`nomeação do presidente` (*nomination for president*)	$(de; nomeação^{\downarrow}, presidente^{\uparrow})$
`nomeou o presidente` (*nominated the president*)	$(dobj; nomear^{\downarrow}, presidente^{\uparrow})$
`discutiu sobre a nomeação` (*disscussed about the nomination*)	$(sobre; discutir^{\downarrow}, nomeação^{\uparrow})$
`nomeação parcial` (*partial nomination*)	$(modif; parcial^{\downarrow}, nomeação^{\uparrow})$

The syntactic context of word w2, $[\lambda x^{\downarrow}(r; x^{\downarrow}, w2^{\uparrow})]$, can be defined extensionally as the set of words that are the *head* of $w2$. The exhaustive enumeration of every word that can occur with that syntactic frame enables us to characterise extensionally its selection restrictions. Similarly, The syntactic context of word $w1$, $[\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$, represents the set of words that are *complement* of $w1$. This set is perceived as the extensional definition of the selection restrictions imposed by the syntactic context. Consider Table 2. The left column contains expressions constituted by two words syntactically related by a particular type of syntactic dependency. The right column contains the syntactic contexts extracted from these expressions. For instance, from theexpression `presidente da república` (*president of the republic*), we extract two syntactic contexts: both $[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$, where `república` plays the role of *complement*, and $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$, where `presidente` is the *head*.

Table 2. Syntactic contexts extracted from binary expressions

Binary Expressions	Syntactic Contexts
`presidente da república` (*president of the republic*)	$[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$, $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$
`nomeação do presidente` (*nomination for president*)	$[\lambda x^{\downarrow}(de; x^{\downarrow}, presidente^{\uparrow})]$, $[\lambda x^{\uparrow}(de; nomeação^{\downarrow}, x^{\uparrow})]$
`nomeou o presidente` (*nominated the president*)	$[\lambda x^{\downarrow}(dobj; x^{\downarrow}, presidente^{\uparrow})]$, $[\lambda x^{\uparrow}(dobj; nomear^{\downarrow}, x^{\uparrow})]$
`discutiu sobre a nomeação` (*disscussed about the nomination*)	$[\lambda x^{\downarrow}(sobre; x^{\downarrow}, nomeação^{\uparrow})]$, $[\lambda x^{\uparrow}(sobre; discutir^{\downarrow}, x^{\uparrow})]$
`nomeação` parcial (*partial nomination*)	$[\lambda x^{\downarrow}(modif; x^{\downarrow}, nomeação^{\uparrow})]$, $[\lambda x^{\uparrow}(modif; parcial^{\downarrow}, x^{\uparrow})]$

Since syntactic configurations impose specific selectional preferences on words, the words that match the semantic preferences (or selection restrictions) required by a syntactic context should constitute a semantically homogeneous word class. Consider the two contexts extracted from `presidente da república`. On the one hand, context $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$ requires a particular noun class, namely human organizations. In corpus P.G.R., this syntactic context selects for nouns such as `república` (*republic*), `governo` (*government*), `instituto` (*institute*), conselho (*council*),... On the other hand, context $[\lambda x^{\downarrow}(r; x^{\downarrow}, república^{\uparrow})]$ requires nouns denoting either human beings or organizations: `presidente` (*president*), `ministro` (*minister of state*), `assembleia` (*assembly*), `governo`, (*government*) `procurador` (*attorney*), `procuradoria-geral` (*attorneyship*) , `ministério` (*state department*), etc.

It follows that the two words related by a syntactic dependency are mutually determined. The context constituted by a word and a specific function imposes semantic conditions on the other word of the dependency. The converse is also true. As has been said, the process of mutual restriction between two related words is called co-specification. In `presidente da república`, the context constituted by the noun `presidente` and the grammatical function *head* somehow restricts the sense of `república`. Conversely, both the noun `república` and the role of *complement* also restrict the sense of `presidente`:

- $[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$ selects for `presidente`
- $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$ selects for `república`

Co-specification is a semantic-syntactic phenomenon which should be taken into account to build distributional word contexts in a more accurate way. Now, in the next subsection, we outline a strategy that defines syntactic context on the basis of simple specification. This results in coarser-grained contexts lacking information which could be usefull for any learning task.

2.5 The Notion of Syntactic Context by Grefenstette

Binary Relations In Grefenstette's strategy [5], syntactic contexts are extracted from binary syntactic dependencies between two words within a noun phrase or between the noun head and the verb head of two related phrases. A binary syntactic dependency could be noted:

$$< r, w1, w2 >$$

where r denotes the syntactic relation itself and w1 and w2 represent two syntactically related words. The syntactic relations used by Grefenstette are: adjective modifiers of nouns (noted ADJ), prepositional modifiers of nouns (NNPREP), nominal modifiers of nouns (NN),[2] verbal subjects (SUBJ), verbal direct objects (DOBJ), and verbal indirect objects (IOBJ). Table 4 displays the the Grefenstette's binary dependencies associated with the same expressions of the previous tables.

Table 3. Binary dependencies by Grefenstette

Related Expressions	Binary Dependencies
`presidente da república` (*president of the republic*)	$< NNPREP, presidente, república >$
`nomeação do presidente` (*nomination for president*)	$< NNPREPnomeaçã, presidente >$
`nomeou o presidente` (*nominated the president*)	$< DOBJ; nomear, presidente >$
`discutiu sobre a nomeação` (*disscussed about the nomination*)	$< IOBJ; discutir, nomeação >$
`nomeação parcial` (*partial nomination*)	$< ADJ; parcial, nomeação^{>}$

Syntactic Contexts Once the binary dependencies have been identified, the system extracts the syntactic contexts. For each word found in the text, the system selects the words that are syntactically related to it. The syntactically related words are considered the syntactic contexts (or attributes) of the given word. In the Grefenstette's approach, special attention is paid to the contexts of nouns. A noun can be syntactically related to an adjective by means of the ADJ relation, to another noun by means of the NN and NNPREP relations, or to a verb by means of SUBJ, DOBJ, and IOBJ relations. These related words are taken to be the known syntactic contexts of the noun. Table **??** shows in the right column the noun contexts that could be extracted provided the binary relations of Table **??**.

In the Grefenstette's notation, the contexts extracted from modifiers of nouns (namely ADJ, and NNPREP modifiers) do not keep the name of the particular syntactic relation. So, <*república*>, is considered as a context of its head noun `presidente`, even though the syntactic relation NNPREP is not explicitly represented. When extracting verbal complements, though, the specific syntactic relation is still available: <*DOBJ*, *nomear*> is a verbal context constituted by both the word related to `presidente` (i.e. the verb `nomear`) and the specific syntactic relation DOBJ.

Note that he notion of syntactic context defined here does not inherit all the available syntactic information from binary dependencies, in particular they do not contain information on the specific preposition relating the two words. This is not semantically appropriate. Take the expressions `discutiu sobre a nominação` (*discused on the nomination*) and `discutiu com o presidente` (*discused with the president*). From these expressions, we extract the same syntactic context $< IDOBJ, discutir >$ for the two nouns: `nominação` and `presidente`. Yet, both nouns should not be considered as having the same syntactic distribution, because they are not related to the verb `discutir` in the same way. In order to formally distinguish the relation beteween

[2] As nominal modifiers of nouns are not common in Portuguese, their meaning is usually syntactically expressed byprepositional phrases.

Table 4. Syntactic Contexts by Grefenstette

Binary Dependencies	Syntactic Contexts of the Head Nouns
$< NNPREP, presidente, república >$ (*president of the republic*)	`presidente`: $< república >$
$< NNPREPnomeaçã, presidente >$ (*nomination for president*)	`nomeação`: $< presidente >$
$< DOBJ; nomear, presidente >$ (*nominated the president*)	`presidente`: $< DOBJ, nomear >$
$< IOBJ; discutir, nomeação >$ (*disscussed about the nomination*)	`nomeação`:$< IOBJ, discutir$
$< ADJ; parcial, nomeação^{>}$ (*partial nomination*)	`nomeação`: $< parcial >$

the verb and the nouns, we must take into account the particular preposition subcategorised by the verb. The preposition would lead us to identify two different syntactic contexts and then two different syntactic distributions of `nominação` and `presidente`.

Nevertheless, the main difference beteween the Grefenstette's strategy and the one presented in the previous section lies on the notion of co-specification.

Syntactic Contexts Defined as Simple Specifications NNPREP relationships are viewd here as *head-complement* dependencies, where only the *head* is specified by the complement. As the specification of the complement by the head is not considered, the head nouns in NNPREP relations cannot be conceived as syntactic contexts of their complements. That is to say, co-specification is not taken into account to characterise syntactic contexts. We claim that simple specification lies on a very conservative conception of syntactic categories. Standard grammar analyses the expression `o presidente da república` (*the president of the republic*) as a relationship between two syntactic categories: the NP (`o presidente` and the PP`da república`. This consevative analysis does not consider the expression `presidente de` as a syntactic constituent at the same level than the PP `da república`. Only less standard grammars, such as Cogntive Grammar ([**?**]), define special grammatical categories for expressions like `presidente de`.[3] We assume that not standard categories represent syntactic contexts at least as semantically significant as the standard categories.

The tests introduced in the following sections attempt to prove that syntactic contexts based on co-specification and, then, containing not standard categories, are more appropriate for acquiring semantic information. Yet, before describing the two learning applications, we will introduce briefly how text corpora is analysed, and how syntactic binary dependencies are identified.

3 Parsing and Identification of Binary Dependencies

The learning techniques were applied on a part of the Portuguese corpus *P.G.R.* (*Portuguese General Attorney Opinions*), which has been previously analysed. This sample is constituted by $1,678,593$ word occurrences, and was analysed in three processing steps. First, it was tagged by the part-of-speech tagger presented in [7]. Then, it was partially analysed by the shallow parser presented in [9]. The shallow parser produced a single partial syntactic description of sentences, which were analysed as sequences of basic chunks (NP, PP, VP, ...). Then, in the third processing step, we used specific attachment heuristics to identify syntactic binary dependencies. Attachment heuristic were based on right association: a chunk tends to attach to another chunk immediately to its right. It was considered that the word heads of two attached chunks form a binary dependency. Binary dependencies were defined and described in the previous section.

It can be easily seen that a great number of syntactic errors may appear since these attachment heuristic does not take into account distant dependencies. Other types of errors are caused, not only by too restrictive attachment heuristics, but also by further misleadings, e.g., words missing from the dictionary, words incorrectly tagged, other sorts of parser limitations, etc. In sum, odd attachments are about 30% over all attachments the system has proposed. None of these errors was manually nor automatically corrected since identification and correction of errors is not a trivial task. Given that any correction on the annoted corpus seems to be not realistic , we decided to apply the learning strategies on noisy text corpora. The semantic information extracted by using these learning strategies should serve in fact to improve the attachment heuristics, and then, to reduce corpus noise in further text processing.

[3] In Cognitive Grammar, `presidente de` and `da república` represent particular instances of the same grammatical category: "Atemporal Relation".

4 Acquisition of Similar Words: The Distributional Hypothesis

The first learning strategy we applied on the *PGR* corpus is to measure word similarity for extracting lists of similar words. Similarity was computed by taking into account the distributional behavior of 4,276 different nouns. Learning is based on the Harry's distributional hypothesis. This section will introduce, first, the particular similarity measure we used to extract lists of similar words. Then, we will make some tests comparing the lists obtained by using informative syntactic contexts (i.e., contexts with information on specific prepositions and co-specification) to the lists obtained from less informative contexts. Finally, we will show a subjective evaluation of these results.

4.1 The Weighted Jaccard Similarity Measure

To compare the syntactic contexts of two words, we used as similarity measure a weighted version of the binary Jaccard measure [5].[4] The binary Jaccard measure, noted BJ, calculates the similarity value between two words, m and n, by comparing the attributes they share and do not share:

$$BJ(word_m, word_n) = \frac{|\{word_m\ attributes \cap word_n\ attributes\}|}{|\{word_m\ attributes \cup word_n\ attributes\}|}$$

The weighted Jaccard measure considers a global and a local weight for each attribute. The global weight gw takes into account how many different words are associated with a given attribute. It is computed by the following formula:

$$gw(attribute_j) = 1 - \sum_i \frac{|p_{ij} \log_2(p_{ij})|}{\log_2(nrels)}$$

where

$$p_{ij} = \frac{frequency\ of\ attribute_j\ with\ word_i}{total\ number\ of\ attributes\ for\ word_i}$$

and $nrels$ is the total number of relations extracted from the corpus. The local weight lw is based on the frequency of the attribute with a given word, and it is calculated by:

$$lw(word_i, attribute_j) = \log_2(frequency\ of\ attribute_j\ with\ word_i)$$

The whole weight w of an attribute is the multiplication of both the global and the local weights. So, the weighted Jaccard similarity WJ between two words m and n is computed by:

$$WJ(word_m, word_n) = \frac{\sum_j \min(w(word_m, attribute_j), w(word_n, attribute_j))}{\sum_j \max(w(word_m, attribute_j), w(word_n, attribute_j))}$$

By computing the similarity measure of all word pairs in the corpus, we extracted the list of the most similar words to each word in the corpus. This process was repeated considering different types of syntactic contexts. On the one hand, we tested the relevance of the use of the prepositional information for the contexts' definition. For this purpose, we compared the results obtained from two types of contexts: "+*prep*–contexts" and "−*prep*–contexts". In the first case, we use syntactic contexts containing information on the specific prepositions, while in the second case we do not use that information. On the other hand, we tested the adequacy of the "↓–contexts" extracted from prepositional dependencies between two noun phrases. For this purpose, we also compared two different types of contexts: "↑↓–'contexts' and "↑–contexts". In the first case, we use contexts with co-specification , while in the second case, we use only contexts with simple specification.

4.2 +*prep*–contexts *versus* −*prep*–contexts

We tested first the contribution of the specific prepositions to measure word similarity. The results obtained from both +*prep*–contexts and −*prep*–contexts, showed that there are no great differences for the words sharing a great number of contexts (namely, more than 100). That is, the results are not significatively different for words frequently appearing in the corpus. Table 5 compares some lists obtained by using +*prep*–contexts to the

Table 5. Similarity lists of frequent words (> 100 different contexts) produced by using contexts with and without prepositional information

Word	Cluster of similar words	
	$+prep$–contexts	$-prep$–contexts
presidente (*president*)	secretário, membro, director, provedor (*secretary, member, director, purveyor*)	director, secretário, procurador, membro (*director, secretary, attorney, member*)
lei (*law*)	artigo, decreto, diploma, norma (*article, decree, diploma, norm*)	artigo, decreto, n, norma (*article, decree, n, norm*)
estado (*state*)	administração, ministério, pessoa, governo (*administration, state department, person, government*)	ministério, administração, governo, tribunal (*state department, administration, government, tribunal*)
diploma (*diploma*)	decreto, lei, artigo, regulamento (*decree, law, article, regulation*)	decreto, lei, artigo, norma (*decree, law, article, norm*)

correspondent lists obtained from $-prep$–contexts, retaining here only words sharing more than 100 contexts. [5]

Nevertheless, when the words sharing less than 100 different contexts (in fact, the most abundant in the corpus) were compared, we observed that the lists obtained from $+prep$–contexts are semanticcally more homogeneous than the lists obtained from $-prep$–contexts. Table 6 shows some of the lists yielded by both types of contexts for less frequently appearing words.

Table 6. Similarity lists of less frequently appearing words (< 100 different contexts) produced by using contexts with and without prepositional information

Word	Lists of similar words	
	$+prep$–contexts	$-prep$–contexts
tempo (*time*)	data, momento, ano, antiguidade (*date, moment, year, seniority*)	década, presidente, admissibilidade, problemática (*decade, president, admissibility, problem*)
regulamento (*regulation*)	estatuto, código, diploma, decreto (*statute, code, diploma, decree*)	membro, decreto, plano, acto (*member, decree, plan, act*)
organismo (*organization*)	autarquia, comunidade, órgão, gabinete (*county, community, organ, cabinet*)	coordenação, dgpc, unidade, ipa (*coordination, dgpc, unit, ipa*)
finalidade (*finality*)	objectivo, escope, fim, objecto (*goal, scope, aim, object*)	capacidade, campo, financiamento, publicidade (*ability, domain, financing, advertising*)
fim (*aim*)	objectivo, finalidade, resultado, efeito (*goal, finality, result, effect*)	decurso, resultado, alvará, apresentação (*duration, result, charter, introduction*)
conceito (*concept*)	noção, regime, estatuto, conteúdo (*notion, regime, statute, content*)	correspondência, grupo, presidente, temática (*correspondence, group, president, subject*)
área (*area*)	ámbito, matéria, processo, sector (*range, matter, process, sector*)	meio, vista, macau, construcção (*mean, view, macau, construction*)

These results deserve special comments. Let's take the lists obtained from the word "tempo" (*time*). The $+prep$–context $[\lambda x^{\uparrow}(de; contrato^{\downarrow}, x^{\uparrow})]$ ($[\lambda x^{\uparrow}(by; contract^{\downarrow}, x^{\uparrow})]$) is shared by "tempo" and "anos" (*years*). As its global weight is quite high (0.78), this context makes the two words more similar. On the contrary, the $-prep$–context $[\lambda x^{\uparrow}(prep; contrato^{\downarrow}, x^{\uparrow})]$ has a very low weight: 0.04. Such a low value makes the attribute not significant when computing the similarity between "tempo" and "anos".

Let's consider another example: the lists obtained from the word "área" (*area*). The $+prep$–context $[\lambda x^{\uparrow}(em; actividade^{\downarrow}, x^{\uparrow})]$ ($[\lambda x^{\uparrow}(in; activity^{\downarrow}, x^{\uparrow})]$) is shared by the words "área", "ámbito" (*range*) and "sector" (*sector*). Given that its global weight is very high (0.94), it contributes to make these three words semantically close. On the contrary, the $-prep$–context $[\lambda x^{\uparrow}(prep; actividade^{\downarrow}, x^{\uparrow})]$ has a lower weight: 0.37. Consequently, it cannot be considered as a significant clue when comparing the similarity between the three words.

[4] We implemented various statistical measures: coefficient of Jaccard, other specific version of the weighted Jaccard, and the particular coefficient of Lin. They did not improve, though, the results obtained from the weighted Jaccard measure described in this section.

[5] We do not use a systematic evaluation methodology based on machine readable dictionaries or electronic thesaurus, because this sort of lexical resources for Portuguese are not available yet.

Therefore, it can be assumed that the information about specific prepositions is relevant to characterise and identify the significant syntactic contexts used for the measurement of word similarity. In the following subsection, we'll show that contexts based on co-specification are at least as significant as contexts with prepositions.

4.3 Co-specification: $x_{\uparrow}$ $x_{\downarrow}$–'contexts' *versus* $x_{\downarrow}$–'contexts'

We also tested the contribution of the $x_{\uparrow}$–contexts (extracted from noun phrases) to yield lists of similar words. These contexts were extracted by taking into account the co-specification hypothesis. The lists obtained from $x_{\uparrow}x_{\downarrow}$–contexts are significantly more accurate than those obtained from simple specification (i.e., from $x_{\downarrow}$–contexts), even for the frequently appearing words such as "diploma" (*diploma*) or "decreto" (*decree*). Table 7 illustrates some of the lists extracted from both types of contexts.

Table 7. Similarity lists produced by contexts with and without $x_{\uparrow}$–contexts

Word	Lists of similar words	
	↑↓–strategy	↑–strategy
juíz (*judge*)	dirigente, presidente, subinspector, governador (*leader*, *president*, *subinspector*, *governor*)	contravenção, vereador, recinto, obrigatoriedade (*contravention*, *councillor*, *enclosure*, *obligatoriness*)
diploma (*diploma*)	decreto, lei, artigo, convenção (*decree*, *law*, *article*, *convention*)	tocante, diploma, magistrado, visita (*concerning*, *diploma*, *magistrate*, *visit*)
decreto (*decree*)	diploma, lei, artigo, n (*diploma*, *law*, *article*, *n*)	ambos, sessão, secretaria, coadjuvação (*both*, *session*, *department*, *cooperation*)
regulamento (*regulation*)	estatuto, código, sistema, decreto (*statute*, *code*, *system*, *decree*)	membro, meio, prejuízo, emissão (*member*, *mean*, *prejudice*, *emission*)
regra (*rule*)	norma, princípio, regime, legislação (*norm*, *principle*, *regime*, *legislation*)	lugar, data, causa, momento (*location*, *date*, *cause*, *moment*)
renda (*income*)	caução, indemnização, reintegração, multa (*guarantee*, *indemnification*, *reimbursement*, *fine*)	fornecimento, instalação, aquisição, construcção (*suply*, *instalation*, *acquisition*, *construction*)
conceito (*concept*)	noção, estatuto, regime, temática (*notion*, *statute*, *regime*, *subject*)	grau, tipicidade, teatro, abordagem (*degree*, *typicality*, *theater*, *approach*)

On the basis of the results illustrated above, it can be assumed that the use of $x_{\uparrow}$–contexts to yield lists of similar words is extremely significant. Indeed, this type of contexts somehow provides information concerning semantic word classes. Consider the $x_{\uparrow}$–contexts $[\lambda x^{\uparrow}(de; capítulo^{\downarrow}, x^{\uparrow})]$ ($[\lambda x^{\uparrow}(of; chapter^{\downarrow}, x^{\uparrow})]$), $[\lambda x^{\uparrow}(de; anexo^{\downarrow}, x^{\uparrow})]$ ($[\lambda x^{\uparrow}(to; attached^{\downarrow}, x^{\uparrow})]$), and $[\lambda x^{\uparrow}(de; conteúdo^{\downarrow}, x^{\uparrow})]$ ($[\lambda x^{\uparrow}(of; content^{\downarrow}, x^{\uparrow})]$), shared by the words "decreto" and "diploma". As those contexts require nouns denoting the same class, namely *documents*, they can be conceived as syntactic patterns imposing the same selectional restrictions to nouns. Consequently, the nouns appearing with those specific $x_{\uparrow}$–contexts should belong to the class of documents.

In the following subsection, we present a way to subjectivelly evaluate the significance of the different types of syntactic contexts to calculate word similarity.

4.4 Subjective Evaluation

Since lexical resources such as machine readable dictionaries or electronic thesauri are not available yet for Portuguese, we cannot implement a systematic way to compare our results to the lists of words appearing in some "golden standard". The only golden standard that can be used to compare the results is the subjective linguistic knowledge of individuals. The subjective evaluation presented in Table 8 is based on the following strategy. First, we implemented two methods for extracting syntactic contexts: the method introducing information on co-specification and specific prepositions into the syntactic contexts (we call it "Co-specification Method"), and the method that does not take into accout such an information for definying contexts ("Grefenstette Method"). Whereas $33,587$ syntactic contexts were extracted by the former method, only $15,420$ contexts were extracted by the latter. Second, for each noun in the corpus, only its most similar noun was selected. We obtained $5,276$ pairs of similar nouns for each method. Among these nouns, we counted those with only one occurrence in the corpus. Then, we filtered the *a priori* best noun pairs for evaluation. We claim that the best pairs fill one of these two conditions: they must have either a similarity measure higher than 0.1, or a number of shared syntactic contexts higher or equal to 10. Note that such a filtering allows us to select both pairs of nouns sharing discriminant syntactic contexts regardless of their number, and pairs of nouns sharing

several syntactic contexts regardeless of their discriminant nature. We filtered 461 noun pairs from the set of pairs obtained by the Co-specification Method (i.e., 8.7%), while we merely filtered 406 noun pairs from those obtained by the Grefenstette Method (i.e., 7.6%). Both groups of filtered noun pairs were, then, evaluated by two different individuals. In particular, the individuals were required to identify the noun pairs that they consider as being semantically homogenous. Individual A considered 90.59% of the Co-Specification pairs as semantic word pairs, against only 82.30% of the Grefenstette pairs. Individual B identified 91.57% semantic pairs out of the Co-specification pairs, against merely 78.04% of the Grefenstette pairs.

Table 8. Evaluation of two word similarity methods

Methods	Extracted Contexts	Number of Pairs	Coverage (%)	Precision (%)	
				Individual A	Individual B
Co-specification Method	33,587	5,276	8.7%	90.59%	91.57%
Grefenstette Method	15,420	5,276	7.6%	82.30%	78.07%

We may infer from this subjective comparison that the contexts based on the co-specification hypothesis have both a larger coverage (8.7%) and more precision ($+-90\%$) than the contexts based on the Grefenstette Method (7.6% of coverage and $+-80\%$ of precision). Note that the former keep a more important coverage than the latter, even though the frequencies of most of the 33,587 co-specification contexts are not statistically significant. By contrast, frequencies of a great part of the 15,420 Grefenstette contexts are quite high and, consequently, the efficiency of these contexts will not improve significantly in larger corpora. This means that coverage and precision will not be modified as far as the text corpora increase. Nevertheless, we make the assumption that co-specification contexts will have at least more coverage in larger text corpora, since most of their contexts need still higher frequencies to achieve efficiency and correctness.

According to these experimental tests, the distributional similarity obtained by co-specification contexts perform better than the similarity calculated by using poorly defined contexts. In the following section, we will show that co-specification contexts are also appropriate to acquire information on selection restrictions.

5 Acquistion of Selection Restrictions: The Contextual Hypothesis

5.1 Contextual Hypothesis

Selection restrictions are the semantic preferences constraining word combination. In most knowledge-poor approaches to learning selection restrictions, the process of inducing and generalising semantic preferences from word cooccurrence frequencies consists in automatically clustering words considered as similar [?,?,?]. As has been said in the previous section, the best-known strategy for measuring word similarity is based on the *distributional hypothesis*, i.e., the words cooccurring in similar syntactic contexts must be clustered into the same semantic class. However, the learning methods based on the distributional hypothesis may arise some shortcomings. More precisely, they may lead to cluster in the same class words that fill different selection restrictions. Let's analyse the following examples taken from [?]:

(a) John worked till late at the *council*
(b) John worked till late at the *office*
(c) the *council* stated that they would raise taxes
(d) the *mayor* stated that he would raise taxes

On the basis of the distributional hypothesis, since *council* behaves similarly to *office* and *mayor* they would be clustered together into the same word class. Nevertheless, the bases for the similarity between *council* and *office* are different from those relating *council* and *mayor*. Whereas *council* shares with *office* syntactic contexts associated mainly with LOCATIONS (e.g., the argument of *work at* in phrases (a) and (b)), *council* shares with *mayor* contexts associated with AGENTS (e.g., the subject of *state* in phrases (c) and (d)). That means that a polysemous word like *council* should be clustered into various semantic word classes, according to its heterogeneous syntactic distribution. Each particular sense of the word is related to a specific type of distribution. Given that the clustering methods based on the distributional hypothesis solely take into account the global distribution of a word, they are not able to separate and acquire its different contextual senses.

In order to extract contextual word classes from the appropriate syntactic constructions, we claim that similar syntactic contexts share the same semantic restrictions on words. Instead of computing word similarity

on the basis of the too coarse-grained distributional hypothesis, we measure the similarity between syntactic contexts in order to identify common selection restrictions. More precisely, we assume that two syntactic contexts occurring with (almost) the same words are similar and, then, impose the same semantic restrictions on those words. That is what we call *contextual hypothesis*. Semantic extraction strategies based on the contextual hypothesis may account for the semantic variance of words in different syntactic contexts. Since these strategies are concerned with the extraction of semantic similarities between syntactic contexts, words will be clustered with regard to their specific syntactic distribution. Such clusters represent context-dependent semantic classes. Except the cooperative system *Asium* introduced in [2,?,?], few or no research on semantic extraction have been based on such a hypothesis.

Likewise in the system *Asium*, we propose a method to learning selection restrictions based on the contextual hypothesis. However, unlike *Asium*, we work on syntactic contexts containing co-specification information. Whereas *Asium* merely uses the subcategorisation information that verbs impose on their nominal complements in the position of Direct or Indirect object, our method also uses the restrictions imposed by the complements on the head verbs, the restrictions by the head nouns on their complements, and the restrictions by the complements on the head nouns. Since co-specification information allows us to extract more significant syntactic contexts, we may enterily automate the learning strategy. The acquisition of semantic preferences is not made cooperatively, as in the *Asium* system, but in an automatic way.

5.2 Methodology

The objective of this learning method is to cluster words in context-dependent semantic classes, which represent the semantic preferences of syntactic contexts. The input is the set of co-specification contexts extracted from the *PGR* corpus. For this learning task, we extracted 211,976 different syntactic contexts. The processes that we implemented to achieve our objective are the following:

Extraction of Contextual Word Sets: For each previously extracted syntactic context, we select its associated set of words. The words appearing in a particular syntactic context form a *contextual word set*. Given that we have 211,976 different syntactic contexts, we extracted 211,976 contextual word sets. Contextual word sets are taken as the input of the processes of filtering and clustering.

Filtering and Clustering: Each contextual word set is statistically compared to the other contextual word sets using a variation of the weighted Jaccard Measure. For each pair of contextual sets considered as similar, we select only the words that they share. The result is a list of semantically homogenous word sets, called *basic classes*. Then, basic classes are successively aggregated by a conceptual clustering method to induce more general classes, which represent extensionally the selection restrictions of syntactic contexts. So, the word clusters obtained by this method represent context-dependent semantic classes. They are the semantic preferences imposed by their associated syntactic contexts.

In the following, we will introduce in a more accurate way the processes of filtering and clustering. Then, we will show and analyse some of the context-dependent classes obtained by our algorrithm.

5.3 Filtering and Clustering

According to the contextual hypothesis introduced above, two syntactic contexts that select for the same words should have the same extensional definition and, then, the same selection restrictions. So, if two contextual word sets are considered as similar, we infer that their associated syntactic contexts are semantic similar and share the same selection restrictions. In addition, we also infer that these contextual word sets are semantically homogeneous and represent a contextually determined class of words. Let's take the two following syntactic contexts and their associated contextual word sets:

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

Since both contexts share a significant number of words, it can be argued that they share the same selection restrictions. Futhermore, it can be inferred that their associated contextual sets represent the same context-dependent semantic class. In our corpus, context $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*to infringe*) is not only considered as similar to context $[\lambda x^{\downarrow}(dobj; violação^{\downarrow}, x^{\uparrow})]$ (*infringement of*), but also to other contexts such as:

- $[\lambda x^{\downarrow}(dobj; respeitar^{\downarrow}, x^{\uparrow})]$ (*to respect*)
- $[\lambda x^{\uparrow}(dobj; aplicar^{\downarrow}, x^{\uparrow})]$ (*to aply*)

In this section, we will specify the procedure for learning context-dependent semantic classes from the previously extracted contextual sets. This will be done in two steps:

- Filtering: word sets are automatically cleaned by removing those words that are not semantically homogenous.
- Conceptual clustering: the previously cleaned sets are successively aggregated into more general clusters. This allows us to build more abstract semantic classes and, then, to induce more general selection restrictions.

Filtering As has been said in the introduction, the cooperative system Asium is also based on the contextual hypothesis [2,?]. This system requires the interactive participation of a language specialist in order to filter and clean the word sets when they are taken as input of the clustering strategy. Such a cooperative method proposes to manually remove from the sets those words that have been incorrectly tagged or analysed. Our strategy, by contrast, intends to automatically remove incorrect words from sets. Automatic filtering consists of the following subtasks:

First, each word set is associated with a list of its most similar 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 their dispersion (global weight) and their relative frequency for each context (local weigth). Word dispersion (global weight) *disp* takes into account how many different contexts are associated with a given word and the word frequency in the corpus. The local weight is calculated by the relative frequency fr of the pair word/context. The weigth of a word with a context $cntx$ is computed by the following formula:

$$W(word_i, cntx_j) = log_2(fr_{ij}) * log_2(disp_i)$$

where

$$fr_{ij} = \frac{frequency\, of\, word_i\, with\, cntx_j}{sum\, of\, frequencies\, of\, words\, occurring\, in\, cntx_j}$$

and

$$disp_i = \frac{\sum_j frequency\, of\, word_i\, with\, cntx_j}{number\, of\, contexts\, with\, word_i}$$

So, the weighted Jaccard similarity WJ between two contexts m and n is computed by[6]:

$$WJ(cntx_m, cntx_n) = \frac{\sum_{common_i}(W(cntx_m, word_i) + W(cntx_n, word_i))}{\sum_j (W(cntx_m, word_j) + W(cntx_n, word_j))}$$

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 sets of words that are not semantically homogenous. Thus, the intersection of two similar sets represents a semantically homogeneous class, which we call *basic class*. Let's take an example. In our corpus, the most similar set to $[\lambda x^{\uparrow}(de; viola\tilde{c}ao^{\downarrow}, x^{\uparrow})]$ (*infringement of*)) is $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*infringe*) . Both sets share the following words:

`sigilo princípios preceito plano norma lei estatuto disposto disposição direito convenção artigo`
(secret principle precept plan norm law statute disposition disposition right convention article)

This basic class does not contain incorret words such as `vez, flagrantemente, obrigação, interesse` (*time, notoriously, obligation, interest*), which were oddly associated to the context $[\lambda x^{\uparrow}(de; viola\tilde{c}ao^{\downarrow}, x^{\uparrow})]$, but which do not appear in context $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$. This class seems to be semantically homogenous because it contains only words referring to legal documents. Once basic classes have been created, they are used by the conceptual clustering algorithm to build more general classes. Note that this strategy do not remove neither infrequent nor very frequent words. Frequent and infrequent words may be semantic significant provided that they occur with similar syntactic contexts.

[6] *common* means that just common words to both contexts m and n are computed

Conceptual Clustering We use an agglomerative (bottom-up) clustering for successivelly 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 [**?**]. These conditions will be discussed below. Figure 1 shows two basic classes associated with two pairs of similar syntactic contexts. $[CONTX_i]$ represents a pair of syntactic contexts sharing the words `preceito, lei, norma` (*precept, law, norm*, and $[CONTX_j]$ represents a pair of syntactic contexts sharing the words `preceito, lei, direito` (*precept, law, right*). Both basic classes are obtained from the filtering process described in the previous section. Figure 2 illustrates how basic classes are aggregated into more general clusters. If two classes fill the conditions that we will define later, 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`. Such a generalisation 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 $[CONTX_i]$.

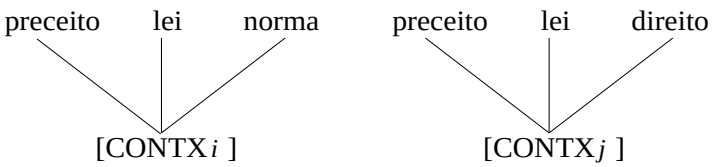

Fig. 1. Basic classes

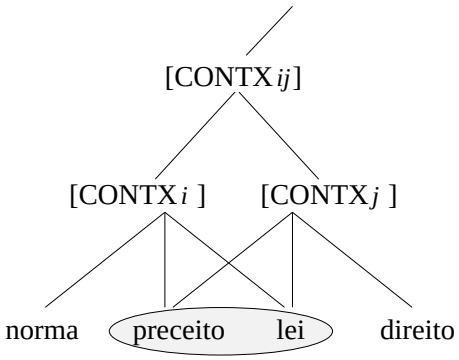

Fig. 2. Agglomerative clustering

Two basic classes are compared and then aggregated into a new more general class if they fulfill three specific conditions:

1. They must have the same number of words. We consider that two classes are compared in a more efficient manner when they have the same number of elements. Indeed, nonsensical results could be obtained if we compare large classes, which still remain polysemic and then heterogenous, to the small classes that are included in them.
2. They must share $n-1$ words. Two classes sharing $n-1$ words are aggregated into a new class of $n+1$ members. Indeed, two classes with the same number of elements only differing in one word may be considered as semantically close.
3. They must have the highest weight. The weight of a class corresponds to the number of occurrences of the class as a subset of other classes (within $n+20$ supersets). Intuitively, the more a class is included in larger classes, the more semantically homogeneous it should be. Only those classes with the highest weight will be compared and aggregated.

Note that clustering does not rely here on a statistical distance between classes. Rather, clustering is guided by a set of constraints, which have been empirically defined considering linguistic data. Due to the nature of these constraints, the clustering process should start with small size classes with n elements, in order to create larger classes of $n+1$ members. All classes of size n that fulfill the conditions stated above are aggregated into $n+1$ clusters. In this agglomerative clustering strategy, level n is defined by the classes with n elements. The algorithm continues merging clusters at more complexe levels and stops when there are no more clusters fulfilling the three conditions.

5.4 Tests and Results

We used a small corpus with 1,643,579 word occurrences, selected from the case-law P.G.R. text corpora. First, the corpus was tagged by the part-of- speech tagger presented in [7]. Then, it was analysed in sequences of basic chunks by the partial parser presented in [9]. The chunks were attached using the right association

heuristic so as to create binary dependencies. 211,976 different syntactic contexts with their associated word sets were extracted from these dependencies. Then, we filter these contextual word sets by using the method described above so as to obtain a list of basic classes.

In order to test our clustering strategy, we start the algorithm with basic classes of size 4 (i.e., classes with 4 elements). We have 7,571 basic classes with 4 elements, but only a small part of them fills the clustering conditions so as to form 1,243 clusters with 5 elements. At level 7, there are still 600 classes filling the clustering conditions, 263 at level 9, 112 at level 11, 38 at level 13, and finally only 1 at level 19. In table 9, we show some of the clusters generated by the algorithm at different intermediate levels.[7]

Table 9. Some clusters at levels 6, 7, 8, 9, 10, and 11

0006 (06)	aludir citar enunciar indicar mencionar referir *allude cite enunciate indicate mention refer*
0009 (07)	considerar constituir criar definir determinar integrar referir *consider constitute create define determinate integrate refer*
0002 (07)	actividade atribuição cargo função funções tarefa trabalho *activity atribution position/task function functions task work*
0003 (08)	administração cargo categoria exercício função lugar regime serviço *administration post rank practice function place regime service*
0002 (09)	abono indemnização multa pensão propina remuneração renda sanção vencimento *bail compensation fine pension fee remuneration rent sanction salary*
0007 (10)	administração autoridade comissão conselho direcção estado governo ministro tribunal órgão *administration autority commission council direction state government minister tribunal organ*
0026 (11)	alínea artigo código decreto diploma disposição estatuto legislação lei norma regulamento *paragraph article code decret diploma disposition statute legislation law norm regulamentation*

Note that some words may appear in different clusters. For instance, `cargo` (*task/post*) is associated with nouns referring to activities (e.g., `actividade, trabalho, tarefa` (*activity, work, task*)), as well as with nouns referring to the positions where those activities are produced (e.g., `cargo, categoria, lugar` (*post, rank, place*)). The sense of polysemic words is represented by the natural attribution of a word to various clusters.

Table 10 illustrates the cluster generated by the algorithm at the last level of aggregation, in this case level 19. This cluster is constituted by words referring to the semantic class of *legal_documents*, which are the more frequent words over the domain-specific corpus P.G.R. Less frequent words are clustered at less high levels. Note that the expressions `n, n1`, or `número` (*number*) do not seem to belong to the class of *legal_documents*. In fact, as they appear in apposition constructions like ``lei number 4'' (*law number 4*), or ``artigo n. 2'' (*law n. 4*), it could be argued that they make reference indirectly to specific legal documents.

Table 10. The cluster generated at level 19

0002 (19)	alínea artigo constituição convenção código decreto diploma disposição estatuto legislação lei n norma n1 número parte preceito regulamento *paragraph article constitution convention code decret diploma disposition statute law n norm n1 number part precept regulamentation*

Since this clustering strategy have been conceived to assure the semantic homogeneity of clusters, it does not really assure that each cluster represents an independent semantic class. Hence, two or more clusters can represent the same contextual-based class and, then, the same semantic restriction. Let's see Table 11. Intuitively, the four clusters, which have been generated at level 12, refer to a general semantic class: *agentive_entities*. Yet, the algorithm is not able to aggregate them into a more general cluster at level 13, since they do not fill condition 2. Further work should refine the algorithm in order to solve such a problem.

Note that the algorithm does not generate ontological classes like *human beings, institutions, vegetables, dogs*,... but context-based semantic classes associated with syntactic contexts. Indeed, the generated clusters

[7] In the left column, the first number represents the weight of the set, i.e., its occurrences as subset of larger supersets; the second number represents class cardinality.

Table 11. Some clusters generated at level 12

0002 (12)	administração associação autoridade comissão conselho direcção entidade estado governo ministro tribunal órgão *administration association autority commission council direction entity state government minister government tribunal organ*
0002 (12)	administração assembleia autoridade comissão conselho direcção director estado governo ministro tribunal órgão *administration assembly autority commission council direction director state government minister government tribunal organ*
0002 (12)	assembleia autoridade câmara comissão direcção estado europol governo ministério pessoa serviço órgão *assembly autority chamber commission direction state europol gouvernment state_department person service organ*
0002 (12)	administração autoridade comissão conselho direcção empresa estado gestão governo ministério serviço órgão *administration autority commission council direction firm state management gouvernment state_department person service organ*

are not linguistic-independent objects but semantic restrictions taking part in the syntactic analysis of sentences. This way, the words `autoridade, pessoa, administração`, etc. (*autority, person, administration*) belong to the same contextual class because they share a great number of syntactic contexts, namely they appear as the subject of verbs such as `aprovar, revogar, considerar`, ... (*aprove, repeal, consider*). Those nouns do not form an ontological class but rather a linguistic class used to constrain the syntactic word combination. More precisely, we may infer that the following syntactic contexts:

$$[\lambda x^{\uparrow}(subj; aprovar^{\downarrow}, x^{\uparrow})]$$
$$[\lambda x^{\uparrow}(subj; revogar^{\downarrow}, x^{\uparrow})]$$
$$[\lambda x^{\uparrow}(subj; considerar^{\downarrow}, x^{\uparrow})]$$

share the same selection restrictions since they are used to build a context-based semantic class constituted by words like `autoridade, pessoa, administração`, ...

In order to evaluate the linguistic significance of the classes acquired by this method , we could use them as semantic heuristics constraining attachment resolution. In that case, we would evaluate the performance of the attachment heuristics. More precisely, if the acquired classes improve the attachment decisions made by the parser, so we can infer that they represent semantic preferences of syntactic contexts. Such an applicative task, however, remains beyond the objectives that limit and circumscribe this paperchapter.

6 Conclusion

In this chapter, we presented the role of the co-specification hypothesis in semantic information acquisition. We argued that syntactic contexts defined on the basis of co-specification make the identification and extraction of semantic information more accurate. Not only they improve word similarity measures based on the distributional strategy, but also they have a suitable performance when used to build context-sensitive classes (as we have described in section sec *contexts*).

Our main aim is to make compatible fine-grained linguistic hypothesis on the complex internal structure of natural langues and unsupervised stochastic strategies such as conceptual clustering. Indeed, only well-defined linguistic features may help us to model the statistic behaviour of words and phrases in an accurate way.

References

1. Crouch, C., Bokyung, Y.: Experiments in automatic statiscal thesaurus construction. 5th Annual International Conference on Research and Development in Information Retrieval, Copenhagen (1992) 77–88
2. Faure, D., Nédellec, C.: ASIUM: Learning subcategorization frames and restrictions of selection. 10th European Conference on Machine Learning, ECML98, Workshop on Text Mining (1998)
3. Ribas Framis, F.: On Learning More Appropriate Selectional Restrictions. 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin (1995)
4. Grefenstette, G.: Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntatic and Window Based Approaches. Corpus processing for Lexical Aquisition, MIT Press, Branimir Boguraev and James Pustejovsky (eds.) (1995) 205–216
5. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publisher (1994)

6. Lin, D.: Automatic Retrieval and Clustering of Similar Words. COLING-ACL'98, Montreal (1998)
7. Marques, N.: Uma Metodologia para a Modelação Estatística da Subcategorização Verbal. Ph.D. Universidade Nova de Lisboa, Portugal (2000)
8. Resnik, P.: Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research **11** (1999) 95–130
9. Rocio, V., de la Clergerie, E., Lopes J.G.: Tabulation for multi-purpose partial parsing. Journal of Grammars **4** (2001)
10. Yarowsky, D.: Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. 14th International Conference on Computational Linguistics, COLING-92, Nantes, France (1992) 454–460
11. Park, Y., Han, Y., Choi K-S.: Automatic thesaurus construction using bayesian networks. International Conference on Information and Knowledge Management, Baltimore (1995) 212–217