USING SYNTAX-BASED METHODS FOR EXTRACTING SEMANTIC INFORMATION

Pablo Gamallo[1]

Caroline Gasperin[2]

Alexandre Agustini[1,2]

José Gabriel Pereira Lopes[1]

Vera Lúcia Strube de Lima[2]

[1] *CITI, Departamento de Informática, Universidade Nova de Lisboa, Portugal*
{gamallo,aagustini,gpl}@di.fct.unl.pt

[2] *PPGCC, Faculdade de Informática, PUCRS, Brasil*
{caroline,vera}@inf.pucrs.br

Abstract This article proposes a sound characterisation of the syntactic features used to acquire semantic information from partially analysed corpora. This characterisation is mainly based on two types of information. First, we take into account the co-specification hypothesis, which states that two syntactically related words impose semantic restrictions to each other. Second, we explore the functional information conveyed, in particular, by prepositions. In order to study the contribution of co-specification and prepositions in different learning tasks, this article describes how the syntactic features defined on the basis of this information can be used to appropriately learn both lists of similar words and classes of selection restrictions. In both cases, we use unsupervised learning strategies.

Keywords: Text Mining, Thesaurus Generation, Selection Restrictions Acquisition.

Introduction

The general aim of this article 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, appropriate, accurate, and well-defined syntactic features in order to acquire sound syntactic-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 (Grefenstette, 1994; Grefenstette, 1995). Knowledge-rich approaches require some sort of previously encoded semantic information (Basili et al., 1993; Framis, 1995; Resnik, 1999): domain-dependent knowledge structures, semantically tagged training corpora, and/or semantic resources such as handcrafted thesauri: Roget's thesaurus, WordNet, and so on. Therefore, knowledge-rich approaches inherits the main shortcomings and limitations of man-made lexical resources. 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 (Pereira et al., 1993; Grefenstette, 1994; Lin, 1998), and selection restrictions (Sekine et al., 1992; Grishman and Sterling, 1994; Dagan et al., 1998). 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: window-based and syntax-based techniques. Window-based techniques consider an arbitrary number of words around a given word window as forming its context. The linguistic information about part-of-speech categories and syntactic groupings is not taken into account in the characterisition of word contexts (Park et al., 1995). The syntax-based strategy, on the contrary, requires specific linguistic information to define word contexts. 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, attachment heuristics are used to specify the possible relations between and within the phrasal groupings. Once this partial 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 regularities in the syntactic distribution of different words (Grefenstette, 1994; Lin, 1998; Faure and Nédellec, 1998).

Both window-based and syntax-based techniques use the Harris' *distributional hypothesis*. According to this assumption, words occurring in similar contexts are considered semantically similar. Usually, the similarity measure between two words is obtained by using their conditional distributions in all contexts. Even though knowledge-poor strategies may differ in the statistical definition of both *conditional distribution* and *similarity measure*, we will not focus on the comparative analysis of these statistical notions for semantic information extraction.

We assume that partial 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, syntax-based approaches can also be significantly different. Different pieces of linguistic information can be taken into account to characterise syntactic contexts. Nevertheless, in the litterature, the choice of a particular type of syntactic context for extracting semantic information is not often properly justified.

This way, the main objective of this article is to establish a specific notion of syntactic context. The appropriateness or the inadequacy of this definition will be tested in two different semantic extraction tasks: word similarity extraction for thesaurus generation and selection restrictions acquisition. And so, this article is organised as follows: In section 1, syntactic contexts will be described on the basis of linguistic co-specification and functional information (prepositions). This notion will be compared to other notions of syntactic contexts. In particular, special attention will be paid to the syntactic contexts used by Grefenstette. Then, in section 3, we will test the appropriateness of our notion of syntactic context compared to other notions, regarding its usefulness 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 achieved by using the Grefenstette's contexts (Grefenstette, 1994). Finally, in section 4, the syntactic contexts we have defined will be used for a different task, namely the acquisition of selection restrictions imposed by words on the words with which they cooccur. It will be claimed that similar syntactic contexts share the same selection restrictions. Our approach will be compared to some elements of the system Asium (Faure and Nédellec, 1998).

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 the learning task easier, given that we have to deal with a limited vocabulary with reduced polysemy.

1. Co-specification and Syntactic Contexts

We argue that the acquisition of linguistic information from corpora can be improved if we take into account 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.

1.1 Co-specification between Predicate and Argument

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 specified. Each word is viewed simultaneously as a predicate imposing restrictions on the words with which it may combine, and as an argument, filling the restrictions imposed by those words.

Consider the relationship between the polysemic verb *load* and the polysemic noun *books* in the non 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”

[1] P.G.R. (*Portuguese General Attorney Opinions*) is constituted by case-law documents in Portuguese (`http://coluna.di.fct.pgr.pt/pgrd/index.html`).

(*rectangular book*), and “symbolic entities” (*naive book*). Yet, the constraints imposed by the words with which it combines allow the noun to be disambiguated. Whereas the adjective *rectangular* activates the physical sense of *book*, the adjective *naive* makes reference to its symbolic content.

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 an unspecified location. The interpretation of the composite expression is not ambiguous any more. Both terms, *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” (Pustejovsky, 1995; Gamallo et al., 2003). 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 simpler operation of argument specification. This operation only permits the one-way specification and disambiguation of the argument by taking into account the sense of the predicate. 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.

1.2 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 dependency relationship between two words (or “binary dependency”), and then, we describe how syntactic contexts are extracted from binary dependencies.

1.2.1 Binary Dependencies. We assume that basic syntactic contexts are extracted from binary syntactic dependencies. Let’s de-

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` (*discussed about the nomination*)	$(sobre; discutir^{\downarrow}, nomeação^{\uparrow})$
`nomeação parcial` (*partial nomination*)	$(modif; parcial^{\downarrow}, nomeação^{\uparrow})$

Table 1. Binary dependencies identified from related expressions

scribe 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, which 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 *dependent* roles, respectively;
- the two words holding the binary relation: $w1$ and $w2$.

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

The binary dependencies (i.e., grammatical relationships) we have considered are the following: subject (noted *subj*), direct object (noted *dobj*), adjective modifier (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 syntactically related words. The right column contains the binary dependencies used to represent these expressions.

1.2.2 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

Binary Dependencies	yntactic Contexts
$(de; presidente^{\downarrow}, república^{\uparrow})$ *(president of the republic)*	$[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$, $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$
$(de; nomeação^{\downarrow}, presidente^{\uparrow})$ *(nomination for president)*	$[\lambda x^{\downarrow}(de; x^{\downarrow}, presidente^{\uparrow})]$, $[\lambda x^{\uparrow}(de; nomeação^{\downarrow}, x^{\uparrow})]$
$(dobj; nomear^{\downarrow}, presidente^{\uparrow})$ *(nominated the president)*	$[\lambda x^{\downarrow}(dobj; x^{\downarrow}, presidente^{\uparrow})]$, $[\lambda x^{\uparrow}(dobj; nomear^{\downarrow}, x^{\uparrow})]$
$(sobre; discutir^{\downarrow}, nomeação^{\uparrow})$ *(discussed about the nomination)*	$[\lambda x^{\downarrow}(sobre; x^{\downarrow}, nomeação^{\uparrow})]$, $[\lambda x^{\uparrow}(sobre; discutir^{\downarrow}, x^{\uparrow})]$
$(modif; parcial^{\downarrow}, nomeação^{\uparrow})$ *(partial nomination)*	$[\lambda x^{\downarrow}(modif; x^{\downarrow}, nomeação^{\uparrow})]$, $[\lambda x^{\uparrow}(modif; parcial^{\downarrow}, x^{\uparrow})]$

Table 2. Syntactic contexts extracted from dependencies

contexts. A syntactic context is extracted by λ-abstracting one of the related words of a binary dependency. Thus, two complementary syntactic contexts can be λ-abstracted from the binary predication associated with a syntactic dependency:

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

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 context enables us to extensionally characterise the selection restrictions imposed by that context. Similarly, the syntactic context of word $w1$, $[\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$, represents the set of words that act as *dependent* of $w1$. This set is perceived as the extensional definition of the selection restrictions imposed by this 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 the expression `presidente da república`, we extract two syntactic contexts: $[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$, where `república` plays the role of *dependent*, and $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$, where `presidente` is the *head*. Here, preposition *de* defines the co-specification relation.

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 *P.G.R.* corpus, this syntactic context selects for nouns such as república (*republic*), governo (*government*), instituto (*institute*), etc. On the other hand, context $[\lambda x^{\downarrow}(r; x^{\downarrow}, \mathit{república}^{\uparrow})]$ requires nouns denoting either human beings or organizations: presidente (*president*), ministro (*minister of state*), assembleia (*assembly*), governo, (*government*) procurador (*attorney*), procuradoria-geral (*general attorneyship*) , ministério (*state department*), etc.

It follows that the two words related by a syntactic dependency are mutually specified. The context defined by a word and a particular 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 noun presidente, the grammatical function *head*, and preposition *de* somehow restricts the sense of república (in other words, context $[\lambda x^{\uparrow}(de; \mathit{presidente}^{\downarrow}, x^{\uparrow})]$ selects for república). Conversely, the noun república, role *dependent*, and preposition *de* also restrict the sense of presidente (i.e., context - $[\lambda x^{\downarrow}(de; x^{\downarrow}, \mathit{república}^{\uparrow})]$ selects for presidente).

Co-specification is a semantic-syntactic phenomenon which should be taken into account to build distributional word contexts in a more accurate way. 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.

1.3 The Notion of Syntactic Context by Grefenstette

1.3.1 Binary Relations. In Grefenstette's strategy (Grefenstette, 1994), 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 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 3 displays in the left column the Grefenstette's binary dependencies associated with the same expressions used in previous tables.

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

Binary Dependencies	Syntactic Contexts of the Head Nouns
< *NNPREP*, *presidente*, *república* > (*president of the republic*)	`presidente:` < *república* >
< *NNPREPnomeação*, *presidente* > (*nomination for president*)	`nomeaçã o:` < *presidente* >
< *DOBJ*, *nomear*, *presidente* > (*nominated the president*)	`presidente:` < *DOBJ*, *nomear* >
< *IOBJ*, *discutir*, *nomeação* > (*discussed about the nomination*)	`nomeaçã o:`< *IOBJ*, *discutir* >
< *ADJ*, *parcial*, *nomeação*> (*partial nomination*)	`nomeaçã o:` < *parcial* >

Table 3. Syntactic Contexts by Grefenstette

1.3.2 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 might be syntactically related to it. Syntactically related words define the syntactic contexts (or attributes) of the given word. In 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 considered syntactic contexts of the noun. Table 3 shows in the right column the noun contexts that could be extracted provided the binary relations of the left column.

In 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`, and the syntactic relation NNPREP is dropped. 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. verb `nomear`) and the specific syntactic relation DOBJ.

Note that the notion of syntactic context used 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. We claim that this does not allow to grasp finer-grained semantic distictions. Take the expressions `discutiu sobre a nominação` (*discussed on the nomination*) and `discutiu com o presidente` (*discused with the president*). From these expressions, we extract the same syntactic context < *IOBJ*, *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 verb discutir (*discuss*) in the same way. In order to formally distinguish the dependeny between a verb and the nouns with which it co-occurs, we must take into account the particular preposition subcategorised by the verb. The preposition leads to the identification of two different syntactic contexts and, then, to two different syntactic distributions of nominação and presidente.

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

1.3.3 Syntactic Contexts Defined as Simple Specifications.

NNPREP relationships are viewed here as *head-dependent* dependencies, where only the *head* is specified by the dependent. As the specification of the *dependent* 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 categorial grammars analyse 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 conservative 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 Cognitive Grammar (Langacker, 1991), define special grammatical categories for complex expressions like presidente de.[3] We assume that non standard categories represent syntactic contexts at least as semantically significant as the standard categories.

Tests introduced in the following sections attempt to show that syntactic contexts based on co-specification 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] In Cognitive Grammar, presidente de and da república represent particular instances of the same grammatical category: "Atemporal Relation".

2. Parsing and Identification of Binary Dependencies

The learning techniques were applied on a part the Portuguese corpus *P.G.R.* (*Portuguese General Attorney Opinions*), which has been previously partially parsed. The training corpus is constituted by $1,678,593$ word occurrences, and was parsed in three processing steps. First, it was tagged by the part-of-speech tagger presented in (Marqes and Lopes, 2001). This tagger reaches 97.3% precision in that corpus. Then, it was partially analysed by the shallow parser presented in (Rocio et al., 2001). The shallow parser produced a single partial syntactic description of sentences, which were analysed as sequences of chunks, i.e., sequences of basic phrases (NP, PP, VP, ...) without dependencies nor recursivity. Then, in the third processing step, we used some specific attachment heuristics to identify syntactic binary dependencies. Attachment heuristics 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.

It can be easily seen that a great number of syntactic errors may appear since these attachment heuristics 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., out of dictionary words, words incorrectly tagged, different types of parser limitations, etc. In sum, odd attachments are about 30% over all attachments the system has proposed. None of these errors was manually or automatically corrected since identification and correction of errors is not a trivial task. Given that any correction on the annotated corpus seems not to be realistic, we decided to apply the learning strategies on noisy text corpora. Semantic information extracted by using these learning strategies is useful to improve the attachment resolution (Gamallo et al., 2003)

3. Acquisition of Similar Words

The aim of this section is to analyse the role of syntactic contexts in the acquisition of lists of similar words. These lists can be further used in applications such as thesaurus generation. Similarity was computed by taking into account the distributional behaviour of $4,276$ different nouns. The learning strategy is based on the Harry's distributional hypothesis. This section will first present the particular similarity measure we used to extract lists of similar words. Then, we will make some tests comparing the lists obtained by using more informative syntactic contexts (i.e., contexts with information on specific prepositions and co-specification) to the lists obtained from less informative ones. Finally, we will show a subjective evaluation of these results.

3.1 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 proposed by (Grefenstette, 1994). The binary Jaccard measure calculates the similarity value between two words by comparing the contexts they share and those they do not share. The weighted Jaccard measure considers a global and a local weight for each context. The global weight gw takes into account the amount of different words associated with a given context. It computes the degree of dispersion of each context by using the following formula:

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

where

$$p_{ij} = \frac{frequency\,of\,cntx_j\,with\,word_i}{total\,number\,of\,contexts\,for\,word_i}$$

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

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

The whole weight w of a context 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:

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

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 used syntactic contexts containing information on specific prepositions, while in the second case we did not use that information. On the other hand, we tested the adequacy of the "$x^{\uparrow}$–contexts" extracted from prepositional dependencies between two noun phrases. For this purpose, we also compared two different types of contexts: "$x^{\uparrow\downarrow}$–contexts" and "$x^{\downarrow}$–contexts". In the first case, we used contexts with co-specification , while in the second case, we only used contexts with simple specification.

3.2 Contribution of Prepositions

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 is no significant difference for words sharing a large number of contexts (namely, more than 100).[4] Nevertheless, when words share less than 100 different contexts (in fact, the most abundant in the corpus), we observed that the lists obtained from $+prep$–contexts are semantically more homogeneous than the lists obtained from $-prep$–contexts. Table 4 shows some of the lists yielded by both types of contexts for less frequently appearing words.

These results deserve special comments. Consider the lists of similar words obtained for noun `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 `ano` (*year*). 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 context not significant when computing the similarity between `tempo` and `ano`.

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

3.3 Contribution of Co-specification

We also tested the contribution of the $x^{\uparrow\downarrow}$–contexts to yield lists of similar words. These contexts were extracted by taking into account the co-specification hypothesis. The lists obtained from $x^{\uparrow\downarrow}$–contexts are significantly more accurate than those obtained from simple specification (i.e., from $x^{\downarrow}$–contexts), even for the frequently occurring words such

[4] 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.

Word	Lists of similar words	
	+*prep*–contexts	−*prep*–contexts
tempo (*time*)	data, momento, ano (*date*, *moment*, *year*)	década, presidente, admissibilidade (*decade*, *president*, *admissibility*)
regulamento (*regulation*)	estatuto, código, decreto (*statute*, *code*, *decree*)	membro, decreto, plano (*member*, *decree*, *plan*)
organismo (*organization*)	autarquia, comunidade, órgão (*county*, *community*, *organ*)	coordenação, dgpc, unidade (*coordination*, *dgpc*, *unit*)
finalidade (*aim*)	objectivo, escope, fim (*goal*, *scope*, *aim*)	capacidade, campo, financiamento (*ability*, *domain*, *funding*)
fim (*aim*)	objectivo, finalidade, resultado (*goal*, *aim*, *result*)	decurso, resultado, alvará (*duration*, *result*, *charter*)
conceito (*concept*)	noção, regime, conteúdo (*notion*, *regime*, *content*)	correspondência, grupo, presidente (*correspondence*, *group*, *president*)
área (*area*)	âmbito, matéria, sector (*range*, *matter*, *sector*)	meio, vista, macao (*mean*, *view*, *macau*)

Table 4. Similarity lists of less frequently appearing words (< 100 different contexts) produced by using contexts with and without prepositional information.

Word	Lists of similar words	
	$x^{\uparrow\downarrow}$–strategy	$x^{\downarrow}$–strategy
juíz (*judge*)	dirigente, presidente, subinspector (*leader*, *president*, *subinspector*)	contravenção, vereador, recinto (*infringement*, *councillor*, *enclosure*)
diploma (*diploma*)	decreto, lei, artigo (*decree*, *law*, *article*)	tocante, diploma, magistrado (*concerning*, *diploma*, *magistrate*)
decreto (*decree*)	diploma, lei, artigo (*diploma*, *law*, *article*)	ambos, sessão, secretaria (*both*, *session*, *department*)
regulamento (*regulation*)	estatuto, código, decreto (*statute*, *code*,*decree*)	membro, meio, prejuízo (*member*, *mean*, *prejudice*)
regra (*rule*)	norma, princípio, regime (*norm*, *principle*, *regime*)	lugar, data, causa (*location*, *date*, *cause*)
renda (*income*)	caução, indemnização, multa (*guarantee*, *indemnification*, *fine*)	fornecimento, instalação, aquisição (*supply*, *installation*, *acquisition*)
conceito (*concept*)	noção, estatuto, temática (*notion*, *statute*, *subject*)	grau, tipicidade, teatro (*degree*, *typicality*, *theatre*)

Table 5. Similarity lists produced by contexts with ($x^{\uparrow\downarrow}$) and without ($x^{\downarrow}$) co-specification.

as `diploma` (*diploma*) or `decreto` (*decree*). Table 5 illustrates some of the lists extracted from both types of contexts.

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 significant. Indeed, this type of contexts somehow provides information concerning semantic word classes. Consider the $x^{\uparrow}$–contexts shared by the words `decreto` and `diploma`: $[\lambda x^{\uparrow}(de; capítulo^{\downarrow}, x^{\uparrow})]$ (*chapter of*), $[\lambda x^{\uparrow}(de; anexo^{\downarrow}, x^{\uparrow})]$ (*annex of*), and $[\lambda x^{\uparrow}(de; conteúdo^{\downarrow}, x^{\uparrow})]$ (*content of*). 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 method to subjectively evaluate the significance of the different types of syntactic contexts to calculate word similarity.

3.4 Subjective Evaluation

Since lexical resources such as machine-readable dictionaries or electronic thesauri are not easily available for Portuguese, we cannot compare our results to the lists of words appearing in some "gold standard". The only standard that can be used to compare the results is the subjective linguistic knowledge of individuals. The subjective evaluation presented in Table 3.4 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 account such an information in the definition of contexts (we call it "Grefenstette Method"). Whereas $33,587$ syntactic contexts sharing at least one word were extracted by the former method, only $15,420$ contexts were extracted by the latter. Second, for each noun in the corpus, only the most similar noun was selected. We obtained $5,276$ pairs of similar nouns for each method. Then, we filtered the *a priori* best noun pairs for evaluation. We assumed that the best pairs must fill one of these two conditions (empirical thresholds): 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 regardless of their discriminant nature. We filtered 461 noun pairs from the set of pairs obtained by the Co-specification Method (i.e., 8.7% coverage), while we merely filtered 406 noun pairs from those obtained by the Grefenstette Method (i.e., 7.6% coverage). 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 considered to be semantically homogeneous. For instance, if the word pair *time-date* was selected, the evaluators are required to check if the two words are somehow semantically related. No specific evaluation criteria have been previously defined. Individual A considered 90.59% Co-Specification pairs as semantically related word pairs, against only 82.30% Grefenstette pairs. Individual B selected 91.57% semantic pairs out of Co-specification pairs, against 78.04% of Grefenstette pairs.

We may infer from this subjective comparison that contexts based on the co-specification hypothesis have both larger coverage (8.7%) and higher precision (≈ 90%) than contexts based on the Grefenstette Method (7.6% coverage and ≈ 80% precision). Note that the former keep a more important coverage than the latter, even though 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, in Grefenstette method, coverage and precision will not be greatly modified as the corpus size grows. By contrast, we make the assumption that co-specification contexts will have at least more coverage in larger text corpora, since most of these contexts still need higher frequencies to achieve efficiency and correctness.

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

Table 6. Evaluation of two word similarity methods

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

4. Acquisition of Selection Restrictions

4.1 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 (Sekine et al., 1992; Grishman and Sterling, 1994; Dagan et al., 1998). As has been said in the previous section, the best-known strategy for measuring word similarity is based on the *distributional hypothesis*, i.e., words cooccurring in similar syntactic contexts must be clustered into the same semantic class. However, learning methods based on the distributional hypothesis may give rise to 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 (Takenobu et al., 1995):

(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. Yet, 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 most similarity methods based on the distributional hypothesis solely take into account the global distribution of a word, they are not able to discriminate its different contextual senses. Some important exceptions are (Pereira et al., 1993; Lin and Pantel, 2001; Allegrini et al., 2000).

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 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 semantically similar. Similar contexts are viewed as contexts imposing the same semantic restrictions. That is what we call *contextual hypothesis.* Semantic extraction strategies based on the contextual hypothesis may account for the semantic variations 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. Few research on semantic extraction has been reported to be based on such a hypothesis. We can cite the cooperative system *Asium* introduced in (Faure and Nédellec, 1998; Faure, 2000), and work by (Reinberger and Daelemans, 2003; Allegrini et al., 2000).

Similarly to 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 dependent nominals (complements) in the position of direct or indirect object, our method also uses the restrictions imposed by the dependent nominals on the head verbs. Since co-specification information allows us to extract more significant syntactic contexts, we may be able to automate to a certain extent the learning strategy. The acquisition of semantic preferences is not made cooperatively, as in the *Asium* system, but automatically

4.2 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 corpus *PGR*. We extracted $211,976$ different syntactic contexts. Then, for each context, we select its associated set of words. 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, which were taken as input for the process of filtering and clustering.

According to the contextual hypothesis introduced above, two syntactic contexts selecting 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 semantically 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:

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

Since both contexts share a significant number of words, it can be argued that they share the same selection restrictions. Furthermore, 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 $[\lambda x^{\downarrow}(de; violação^{\downarrow}, x^{\uparrow})]$ (*infringement of*), but also to other semantically related contexts such as: $[\lambda x^{\downarrow}(dobj; respeitar^{\downarrow}, x^{\uparrow})]$ (*to respect*) and $[\lambda x^{\uparrow}(dobj; aplicar^{\downarrow}, x^{\uparrow})]$ (*to apply*).

In the following, 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: 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.

4.3 Filtering

As has been said, the cooperative system, Asium, is also based on the contextual hypothesis (Faure and Nédellec, 1998; Faure, 2000). This system requires the interactive participation of a language specialist in order to clean the word sets used in the clustering process. 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 Lin coefficient (Lin, 1998), where words are weighted considering their dispersion (global weight) and their relative frequency for each context (local weight). 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 weight of a word with a context 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 Lin similarity lin between two contexts m and n is computed by[5]:

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

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 can 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 $[\lambda x^{\uparrow}(de; violação^{\downarrow}, x^{\uparrow})]$ (*infringement of*)) is $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*infringe*) . Both sets share the following words:

> `princípios preceito plano norma lei estatuto direito artigo`
> *(principle precept plan norm law statute right article)*

This basic class does not contain incorrect words such as `flagrantemente, vez obrigação, interesse` (*notoriously, time, obligation, interest*), which were oddly associated to context $[\lambda x^{\uparrow}(de; violação^{\downarrow}, x^{\uparrow})]$, 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 does not remove neither infrequent nor very frequent words. Frequent and infrequent words may be semantic significant provided that they occur with similar syntactic contexts.

4.4 Conceptual Clustering

We use an agglomerative (bottom-up) clustering for successively aggregating the previously created basic classes. Unlike most research on

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

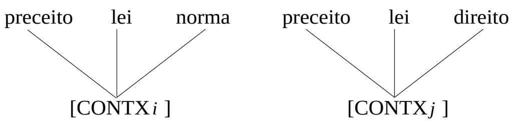

Figure 1. Basic classes

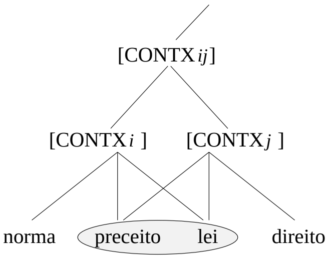

Figure 2. Cluster classes

conceptual clustering, aggregation does not rely only on a statistical distance between classes, but on empirical conditions and constraints (Talavera and Béjar, 1999). 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 $[CONTX_i]$. Two basic classes are compared and then aggregated into a new more general class if they fulfil three specific conditions:

1 They must have the same number n 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 heterogeneous, 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 is driven 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 fulfil 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 complex levels and stops when there are no more clusters fulfilling the conditions. More traditional agglomerative clustering techniques were tested and the type of associations obtained did not seem reasonable. The work by Faure and Naudéllec overtakes these problems using a collaborative technique.

4.5 Tests and Results

We extracted 211,976 different syntactic contexts with their associated word sets from *P.G.R.* text corpora. Then, we filter these contextual word sets by using the method described above in order 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 7, we show some of the clusters generated by the algorithm at different intermediate levels.[6]

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

[6] 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.

006 (06)	`aludir citar enunciar indicar mencionar referir` *allude cite enunciate indicate mention refer*
009 (07)	`considerar constituir criar definir determinar integrar referir` *consider constitute create define determinate integrate refer*
002 (07)	`actividade atribuição cargo função funções tarefa trabalho` *activity attribution position/task function functions task work*
003 (08)	`administração cargo categoria exercício função lugar regime serviço` *administration post rank practice function place regime service*
002 (09)	`abono indemnização multa pensão propina remuneração renda sanção vencimento` *bail compensation fine pension fee remuneration rent sanction salary*
007 (10)	`cámara comissão direcção estado europol governo ministério pessoa serviço órgão` *city_corporation commission direction state europol government state_department person service organ*
017 (14)	`alínea artigo código convenção decreto diploma disposição estatuto legislação lei norma n regime regulamento` *paragraph article code convention decree certificate disposition statute legislation law norm n regime regulation*

Table 7. Clusters at different levels

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 assignment of a word to various clusters.

Note as well 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 are not linguistic-independent objects but semantic restrictions taking part in the syntactic analysis of sentences. This way, the words `direcção, pessoa, estado`, etc. (*direction, person, state*) 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`, ... (*approve, repeal, consider*). Those nouns do not form an ontological class but rather a linguistic class used to constrain the syntactic word combination. So, we may infer that contexts like $[\lambda x^{\uparrow}(subj; aprovar^{\downarrow}, x^{\uparrow})]$ and $[\lambda x^{\uparrow}(subj; revogar^{\downarrow}, x^{\uparrow})]$ share the same selection restrictions since they are used to build a context-based semantic class constituted by words like `direcção, pessoa, estado`, etc. By contrast, ontological classes (i.e., *vegetables*) are rarely used to characterise the selection restrictions of a set of similar syntactic contexts.

In order to evaluate the linguistic significance of the classes acquired by this method , we are using them as semantic heuristics constraining attachment resolution. In that case, we will evaluate the performance of the attachment heuristics. More precisely, if the acquired classes improve the attachment decisions made by a parser, so we can infer that they represent semantic preferences of syntactic contexts. Such an applicative task remains beyond the objectives that limit and circumscribe this article. Details of this syntactic evaluation can be seen in (Gamallo et al., 2003).

5. Summary

In this article, we analysed the role of a particular notion of syntactic context in semantic information acquisition. In particular, we describe the semantic behaviour of two linguistic components of contexts: both co-specification and prepositional information. We argued that syntactic contexts defined on the basis of co-specification and specific prepositions 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. Concerning the latter task, we make the assumption that similar syntactic contexts share the same selection restrictions and then requires similar context-sensitive classes. In order to learn these classes, we account for a particular notion of linguistic similarity: we measure, not similarity between words on the basis of their syntactic distribution, but similarity between syntactic contexts on the basis on the word distribution (as we have described in section 4).

The main aim of the article was to make compatible fine-grained linguistic hypothesis on the structure of natural languages (like co-specification) 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.

In current work, we are using the thesaurus of similar words as a lexical resource constraining the way we built context-sensitive classes. So, we integrate the results of our first task (described in section 3) into the clustering process described in section 4. The new classes obtained by this extended technique are being evaluated by measuring their performance in several NLP applications: attachment resolution, word sense disambiguation, and information retrieval.

Acknowledgments

The authors would like to thank their sponsors: POSI programme, Fundação para a Ciência e a Tecnologia, Portugal; Federal Agency for

Post-Graduate Education (CAPES), Brazil; Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil.

References

Allegrini, P., Montemagni, S., and Pirrelli, V. (2000). Learning word clusters from data types. In *Coling-2000*, pages 8–14.

Basili, R., Pazienza, M., and Velardi, P. (1993). Hierarchical clustering of verbs. In *Workshop on Acquisition of Lexical Knowledge from Text*, pages 56–70, Ohio State University, USA.

Dagan, I., Lee, L., and Pereira, F. (1998). Similarity-based methods of word coocurrence probabilities. *Machine Learning*, 43.

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

Faure, D. and Nédellec, C. (1998). Asium: Learning subcategorization frames and restrictions of selection. In *ECML98, Workshop on Text Mining.*

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

Gamallo, P., Agustini, A., and Lopes, G. P. (2003). Learning subcategorisation information to model a grammar with co-restrictions. *Traitement Automatic de la Langue*, 44(1):93–117.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, USA.

Grefenstette, G. (1995). Evaluation techniques for automatic semantic extraction: Comparing syntatic and window based approaches. In Boguraev, B. and Pustejovsky, J., editors, *Corpus processing for Lexical Acquisition*, pages 205–216. The MIT Press.

Grishman, R. and Sterling, J. (1994). Generalizing automatically generated selectional patterns. In *Proceedings of the 15th International on Computational Linguistics (COLING-94).*

Langacker, R. W. (1991). *Foundations of Cognitive Grammar: Descriptive Applications*, volume 2. Stanford University Press, Stanford.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal.

Lin, D. and Pantel, P. (2001). Induction of semantic classes from natural language text. In *SIGKDD-01*, Montreal, Canada.

Marqes, N. and Lopes, G. (2001). Tagging with small training corpora. In Hoffmann, F., Hand, D., Adams, N., Fisher, D., and Guimaraes, G., editors, *Advances in Intelligent Data Analysis*, pages 62–72. LNCS, Springer Verlag.

Park, Y., Han, Y., and Choi, K.-S. (1995). Automatic thesaurus construction using bayesian networks. In *International Conference on Information and Knowledge Management*, pages 212–217, Baltimore.

Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Association of Comptutational Linguistics*, pages 183–190, Columbos, Ohio.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge.

Reinberger, M.-L. and Daelemans, W. (2003). Is shallow parsing useful for unsupervised learning of semantic clusters? In *4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-03)*, pages 304–312, Mexico City.

Resnik, P. (1999). Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

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

Sekine, S., Carrol, J., Ananiadou, S., and Tsujii, J. (1992). Automatic learning for semantic collocation. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 104–110.

Takenobu, T., Makoto, I., and Hozumi, T. (1995). Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI-95*.

Talavera, L. and Béjar, J. (1999). Integrating declarative knowledge in hierarchical clustering tasks. In *Intelligent Data Analysis*, pages 211–222.