Syntactic-Based Methods for Measuring Word Similarity

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 paper explores different strategies for extracting similarity relations between words from partially parsed text corpora. The strategies we have analysed do not require supervised training nor semantic information available from general lexical resources. They differ in the amount and the quality of the syntactic contexts against which words are compared. The paper presents in details the notion of syntactic context and how syntactic information could be used to extract semantic regularities of word sequences. Finally, experimental tests with Portuguese corpus demonstrate that similarity measures based on fine-grained and elaborate syntactic contexts perform better than those based on poorly defined contexts.

1 Introduction

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 [4]. Knowledge-rich approaches require some sort of previously encoded semantic information [8, 3, 10]: 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 attempt to extract the frequency of co-occurrences of words within various contexts to compute semantic similarity among words. More precisely, the similarity measure takes into account the contexts that words share or do not share, as well as the importance of these contexts for each word. Words which share a great number of contexts are considered as similar.

Since contexts can be defined in two different ways, 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

* 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.

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 category to each word of the corpus. Then, the tagged corpus is segmented into a sequence of 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. Then, a statistical method compares the frequency of the shared contexts to judge word similarity [5, 6, 2]. In both strategies, window-based and syntactic-based techniques, words will be compared to each other in terms of their contexts; yet, we consider that syntactic analysis opens up a much wider range of more precise 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[1]. 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 measuring word similarity has not been properly justified by those researchers.

This way, the main objective of this paper is to analyse the appropriateness or the inadequacy of different types of syntactic contexts for computing word similarity. More precisely, various syntactic-based strategies will be compared on the basis of different definitions of the notion of syntactic context. For this purpose, we apply these strategies on the Portuguese corpus *P.G.R.*[2] The article is organised as follows. In the next section, various types of syntactic contexts will be analised. Special attention will be paid to the notion of syntactic context used by Grefenstette, as well as for the specific notion that we have defined. Then, in section 3.1, we will use the same statistical similarity measure to compare the appropriateness of the syntactic contexts defined in the previous section. The best results are obtained when the syntactic-based strategy relies on our notion of syntactic context. Samples of the results we have obtained are presented in the Appendix.

2 Types of Syntactic Contexts

In this section, we analyse the notion of syntactic context used by Grefenstette to compute similar words [5]. Then, we extract further syntactic information from the partially parsed text in order to make syntactic contexts more elaborate. As a consequence, we obtain fine-grained contexts which contain more specific information than the one provided by Grefenstette's approach.

2.1 The Notion of Attribute by Grefenstette

Grefenstette calls "attributes" the syntactic contexts of a word. Attributes 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:

$$< R, w1, w2 >$$

[1] The choice of a particular measure of similarity may be another parameter to compare various syntactic-based approaches

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

where R denotes the syntactic relation itself (e.g., ADJ, NN, NNPREP, SUBJ, DOBJ, and IOBJ), and w1 and w2 represent two syntactically related words. Table 1 shows some syntactic dependencies between the noun "cause" and other related words.

Expressions	Binary Dependencies
`possible causes`	<*ADJ*, *cause*, *possible*>
`the cause of neonatal jaundice`	<*NNPREP*, *cause*, *jaundice*>
`no cause could be determined`	<*DOBJ*, *determine*, *cause*>
`death cause`	<*NN*, *cause*, *death*>

Table 1. Exemples of syntactic dependencies

Then, 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 attributes of the given word, i.e., its syntactic contexts. For instance, 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 attributes of the noun.

In order to select the attributes of "cause", the system takes as input all the binary dependencies between "cause" and other words. Then, it extracts all the specific words syntactically related to "cause", since they represent its particular attributes. For example, from the 4 dependencies illustrated in 1 between "cause" and another word, it is possible to extract 4 attributes of "cause" (see table 2).

Binary Dependencies	Attributes of `cause`
<*ADJ*, *cause*, *possible*>	<*possible*>
<*NNPREP*, *cause*, *jaundice*>	<*jaundice*>
<*DOBJ*, *determine*, *cause*>	<*DOBJ*, *determine*>
<*NN*, *cause*, *death*>	<*death*>

Table 2. Attributes of `cause`

In the Grefenstette's notation, the attributes extracted from noun modifiers (namely NN, ADJ, and NNPREP modifiers) do not keep the name of the particular syntactic relation. So, <*jaundice*>, <*possible*>, and <*death*> are attributes of "cause" even though the syntactic relations NNPREP, ADJ and NN are not explicitly represented. When extracting verbal complements, though, the specific syntactic relation is still available: <*DOBJ*, *determine*> is a verbal attribute constituted by both the word related to "cause" (i.e. the verb "determine") and the specific syntactic relation DOBJ.

2.2 Underspecified Attributes

The notion of attribute defined in the previous section does not inherit all the available syntactic information from binary dependencies. Consider one of the Portuguese expressions found in our corpus: "autorização à empresa" (*permission to the company*). From this expression, <*empresa*> (*company*) is extracted as the attribute of "autorização" (*permission*). Yet, relevant information implicitly contained in the dependency relation has been lost:

– information about the specific preposition: the attribute <*empresa*> does not convey information about the particular preposition “a” relating the two words ;
– information about the opposite attribute: the attribute <*autorização*> modifying the word “empresa” is not considered.

Information about prepositions should be taken into account since they convey important syntactic and semantic information. Let’s consider two prepositional expressions: “autorização à empresa” (*permission to the company*) and “ autorização da empresa” (*authorization by the company*). According to the Grefenstette’s notion of attribute, we should extract the same attribute, namely <*empresa*>, from both expressions. Nevertheless, preposition “a” (*to*) introduces a quite different syntactic dependency than the one introduced by preposition “de” (*by*). Whereas the preposition “a” requires <*empresa*> to be the receiver within the action of giving authorazation, the preposition “de” requires <*empresa*> to be the agent of this action. Therefore, for the purpose of extracting semantic regularities, prepositions should be considered as internal facets of syntactic contexts.

¿From the Grefenstette’s viewpoint, only one attribute, <*empresa*>, could be extracted from the NNPREP expression “ autorização à empresa”. Whereas the modifier word (i.e., the noun after the preposition) is considered as a potential attribute, the modified word (i.e., the noun before the preposition) cannot become an attribute. The tests introduced in section 3.1 will show that the Grefenstette’s notion of attribute is too restrictive for the purpose of measuring word similarity.

2.3 More Accurate Attributes for Word Similarity Measurement

In order to take into account the implicit information contained in the dependency relationships, we will introduce a more general and flexible definition of attribute. The results of the computational tests presented in the next section will provide us with empirical evidence about the appropriateness of such a definition.

Attributes are extracted from binary syntactic dependencies. A syntactic dependency may be represented 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 *modified* and *modifier* roles, respectively ;
– the two words holding the binary relation: w1 and w2.

In this binary syntactic dependency, the word indexed by “$^{\downarrow}$” plays the role of *modified*, whereas the word indexed by “$^{\uparrow}$” plays the role of *modifier*. Therefore, w1 is modified by w2 as well as w2 modifies w1. This way, two complementary attributes may be extracted from that syntactic dependency:

$$< \downarrow r, w1 >< \uparrow r, w2 >$$

where $< \downarrow r, w1 >$ is the attribute of w2 and $< \uparrow r, w2 >$ is the attribute of w1. An attribute is defined as the pair constituted by both a specific syntactic function and the word associated to this function. In particular, $\downarrow r$ represents the syntactic function of *modified*, and $\uparrow r$ the *modifier* function. Consider Table 3. The left column contains expressions constituted by two words syntactically related by a particular type of syntactic dependency. The right column contains the attributes extracted from these expressions. For instance, from the expression "autorização á empresa", it was extracted both the attribute $<\uparrow a,$ *empresa*$>$, where "empresa" plays the role of modifier word, and the attribute $<\downarrow a,$ *autorização*$>$, where "autorização" is the modified word. Furthermore, information about the specific preposition connecting the words is also available. Our notion of attribute is closely related to what Lin calls "feature" [6].

Binary Expressions	Attributes
`autorização à empresa` (*permission to the company*)	$<\uparrow a,$ *empresa*$>$, $<\downarrow a,$ *autorização*$>$
`nomeação do presidente` (*appointment of the president*)	$<\uparrow de,$ *presidente*$>$, $<\downarrow de,$ *nomeação*$>$
`nomeou o presidente` (*appointed the president*)	$<\uparrow dobj,$ *presidente*$>$, $<\downarrow dobj$ nomear$>$
`discutiu sobre a nomeação` (*disscussed about the appointment*)	$<\uparrow sobre,$ *nomeação*$>$, $<\downarrow sobre,$ discutir$>$

Table 3. Elaborate attributes

These elaborate attributes provide us with fine-grained syntactic contexts. In the following section, we will compare these informative syntactic contexts to the coarse-grained contexts used by Grefenstette. This will lead us to assume that the elaborate information conveyed by our notion of attribute is able to contribute more accurately to design a suitable strategy for clustering similar words.

3 Comparing Syntactic-Based Strategies

Various semantic extraction techniques were applied to the Portuguese corpus P.G.R. (*Portuguese General Attorney Opinions*), which is constituted by 1,678,593 word occurrences. These techniques use the output obtained from two previous pre-processing steps. First, the corpus was tagged by the part-of- speech tagger presented in [7]. Then, it was analysed by the partial parser presented in [9]. Similarity was computed by measuring the syntactic information shared by 4,276 different nouns on the basis of 186,952 different attributes.

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 [5].[3] 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:

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

[3] We implemented various statistical measures: coefficient of Jaccard, a 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.

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 attributes' definition. For this purpose, we compared the results obtained from two strategies: "+*prep*–strategy" and "−*prep*–strategy". The former uses attributes containing information about the specific prepositions, while the latter does not use that information. On the other hand, we tested the adequacy of the "↓–attributes" extracted from prepositional dependencies between two noun phrases. For this purpose, we also compared two different methods: "↑↓–strategy" and "↑–strategy". The former contains both types of attributes, while the later uses only ↑–attributes.

3.2 +*prep*–strategy *versus* −*prep*–strategy

We tested first the contribution of the specific prepositions to measure word similarity. The results obtained from both strategies, +*prep*– strategy and −*prep*–strategy, showed that there are no great differences for the words sharing a great number of attributes (namely, more than 100 attributes). That is, the results are not significatively different for words frequently appearing in the corpus. Table 4 compares the results obtained from the +*prep*- -strategy to those obtained from the −*prep*–strategy, retaining at random some of the words sharing more than 100 attributes.[4]

Nevertheless, when the words sharing less than 100 attributes (in fact, the most abundant in the corpus) was compared, we observed that the lists obtained from the +*prep*–strategy are semanticcally more homogeneous than the lists obtained from the −*prep*–strategy. Table 5 shows some of the lists yielded by these strategies for less frequently appearing words.

These results deserve special comments. Let's take the lists obtained from the word "tempo" (*time*). In the +*prep*–strategy, the attribute <↓*por*, *contrato*> (<↓*by*, *contract*>) is shared by

[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	Cluster of similar words	
	+*prep*–strategy	−*prep*–strategy
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*)

Table 4. Similarity lists of frequently appearing words (> 100 attributes) produced by contexts with and without prepositional information

Word	Cluster of similar words	
	+*prep*–strategy	−*prep*–strategy
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*)	âmbito, matéria, processo, sector (*range*, *matter*, *process*, *sector*)	meio, vista, macau, construcção (*mean*, *view*, *macau*, *construction*)

Table 5. Similarity lists of less frequently appearing words (< 100 attributes) produced by contexts with and without prepositional information

"tempo" and "anos" (*years*). As its global weight is quite high (0, 78), this attribute make the two words more similar. On the contrary, in the −*prep*–strategy, the no–prepositional attribute <↓, *contrato*> 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*). In the +*prep*–strategy, the attribute <↓*em*, *actividade*> (<↓*in*, *activi*ty>) is shared by the words "área", "âmbito" (*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, in the −*prep*–strategy, the no–prepositional attribute <↓, *actividade*> has a lower weight: 0, 37. Consequently, it cannot be considered as a significant clue when comparing the similarity between the three words.

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.

3.3 $\uparrow\downarrow$–strategy *versus* $\uparrow$–strategy

We also tested the contribution of the $\downarrow$–attributes (extracted from noun phrases) to yield lists of similar words. The lists obtained from $\uparrow\downarrow$–strategy are significantly more accurate than those obtained from the $\uparrow$–strategy, even for the frequently appearing words such as "diploma" (*diploma*) or "decreto" (*decree*). Table 6 illustrates some of the lists extracted from both strategies.

Word	Cluster of similar words	
	$\uparrow\downarrow$–strategy	$\uparrow$–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*)

Table 6. Similarity lists produced by contexts with and without $\downarrow$–attributes

On the basis of the results illustrated above, it can be assumed that the use of $\downarrow$– attributes to yield lists of similar words is extremely significant. Indeed, this type of attributes somehow provides information concerning the semantic word class. Consider the $\downarrow$–attributes $<\downarrow de$, *capítulo*$>$ ($<\downarrow of$, *chapter*$>$), $<\downarrow a$, *anexo*$>$ ($<\downarrow to$, *attatched*$>$) and $<\downarrow de$, *conteúdo*$>$ ($<\downarrow of$, *content*$>$), shared by the words "decreto" and "diploma". As those attributes 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 $\downarrow$–attributes should belong to the class of documents.

In the Appendix, we compare the lists extracted by using the fine-grained techniques (i.e., both $\uparrow\downarrow$–strategy and +*prep*–strategy) to the lists extracted by using the coarse-grained methods: $\uparrow$–strategy and −*prep*–strategy. The words constituting the lists obtained from the more informative strategies are semantically more homogeneous than those obtained from the less informative ones.

4 Final Remarks

According to the results of the tests described above, the strategies based on rich syntactic contexts are more accurate for the measurement of similar words. Indeed, the specific syntactic data that we used to refine syntactic attributes allows us to identify more informative syntactic-semantic dependencies between words. Experimental tests demonstrated that similarity measures relied on the fine-grained syntactic contexts we have defined in this paper perform better than those based on poorly defined contexts.

Nevertheless, all the syntactic-based approaches (fine-grained and coarse- grained syntactic strategies) are confronted with two sorts of linguistic phenomena: both polysemic words and odd attachments of syntactic dependencies between words.

The lists of words recognised as being similar to a particular word can be semantically heterogeneous because of the lexical polysemia of the compared word. For example, the word "contrato" (*contract*) appears to be similar to words describing activities such as "trabalho" (*work*), "acto" (*act*), "processo" (*process*), as well as to words describing documents such as "norma", "regulamento". It shares with the former group attributes requiring a sort of activity (e.g., $<\downarrow de$, *execução*$>$ ($<\downarrow of$, *execution*$>$), : $<\downarrow de$, *conclusão*$>$ ($<\downarrow of$, *end*$>$)), and shares with the later group attributes requiring documents: $<\downarrow dobj$, *assinar*$>$ ($<\downarrow dobj$, *sign*$>$). Various clustering proposals have been made so as to group the similar words together along sense axes [5, 2]. However, the implementation of a efficient clustering method used for the groupping of semantically homogenous words remains to be future work.

Finally, attachment errors inherited from the parser should be taken into account. The number of these errors increases when the language analysed is not as syntactically rigid as English. When considering our Portuguese corpus, we found that of a sample of analysed text more than 30% were incorrect attachments. This weak rate is probably due to the non predictable constraints on the Portuguese syntactic order. To palliate the noisy information inherited from the parser limitations for Portuguese, we are impelled to find more robust word similarity strategies than those used for English texts. So, the fine-grained strategies defined in this paper could be perceived as an attempt to partially palliate the poor results obtained by parsing Portuguese text corpora.

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

Appendix

Word	Cluster of similar words	
	+*prep* –strategy and ↑↓–strategy	−*prep*–strategy and ↑–strategy
vontade (*will*)	pensamento, iniciativa, comportamento, designação (*thought*, *initiative*, *behavior*, *designation*)	postura, conformação, número, impedimento (*posture*, *conformation*, *number*, *impediment*)
trabalho (*job*)	serviço, actividade, informação, função (*service*, *activity*, *information*, *function*)	promoção, subsídio, exercício, programa (*promotion*, *subsidy*, *exercise*, *program*)
tempo (*time*)	ano, período, data, prazo (*year*, *period*, *date*, *delay*)	liberdade, inscrição, montante, momento (*liberty*, *inscription*, *amount*, *moment*)
sentença (*sentence*)	condenação, decisão, pena, custa (*condemnation*, *decision*, *punishment*, *cost*)	moldura, ilícito, decisão (*frame*, *illicit*, *decision*)
secretaria (*department*)	secção, provedoria, ministério, gabinete (*section*, *purveyor's office*, *state department*, *cabinet*)	correspondente, dgpe, escrivão, cj (*correspondent*, *dgpe*, *clerk*, *cj*)
renda (*income*)	indemnização, caução, quota, multa (*indemnification*, *guarantee*, *quota*, *fine*)	fornecimento, instalação, libertação, indemnização (*suply*, *instalation*, *liberation*, *indemnification*)
regra (*rule*)	norma, princípio, regime, disposição (*norm*, *principle*, *regime*, *disposition*)	promoção, obrigatoriedade, directiva, autarquia (*promotion*, *obligatoriness*, *directive*, *autarchy*)
prazo (*delay*)	limite, período, condição, termo (*limit*, *period*, *condition*, *term*)	data, deliberação, obrigatoriedade, alvará (*date*, *deliberation*, *obligatoriness*, *charter*)
pagamento (*payment*)	montante, cobrança, cumprimento, despesa (*amount*, *exaction*, *accomplishment*, *expense*)	montante, cobrança, fixação, juro (*amount*, *expence*, *fixation*, *interest*)
organismo (*organization*)	órgão, autarquia, instituição, departamento (*organ*, *autarchy*, *institution*, *department*)	militar, magistrado, deliberação, fracção (*soldier*, *magistrate*, *deliberation*, *fraction*)
objectivo (*goal*)	finalidade, fim, escopo, objecto (*finality*, *aim*, *scope*, *object*)	tarefa, núcleo, resultado, político (*task*, *center*, *result*, *politician*)
lei (*law*)	diploma, norma, artigo, decreto (*diploma*, *norm*, *article*, *decree*)	diploma, disposição, decreto, norma (*diploma*, *disposition*, *decree*, *norm*)
investigador (*investigator*)	assessor, inspector, auxiliar, subdirector (*assessor*, *inspector*, *assistant*, *subdirector*)	lugar, finalidade, vencimento, recrutamento (*position*, *finality*, *expiration*, *recruiting*)
governo (*government*)	ministro, governador, assembléia, tribunal (*minister of state*, *governor*, *assembly*, *tribunal*)	governador, ministro, assembléia, tribunal (*governor*, *minister of state*, *assembly*, *tribunal*)
fim (*aim*)	finalidade, objectivo, objecto, efeito (*finality*, *goal*, *object*, *effect*)	intuito, decurso, autarquia, recuperação (*intention*, *duration*, *autarchy*, *recuperation*)
estado (*state*)	administração, pessoa, entidade, ministério (*administration*, *person*, *entity*, *state department*)	membro, ministro, tribunal, poderão (*member*, *minister of state*, *tribunal*)
director (*director*)	presidente, governador, ministro, provedor (*president*, *governor*, *minister of state*, *purveyor*)	provedor, craf, presidente, governador (*purveyor*, *craf*, *president*, *governor*)
despacho (dispatch)	portaria, decreto, decisão, proposto (*order*, *decree*, *decision*, *proposed*)	gabinete, ambos, portaria, sessão (*cabinet*, *both*, *order*, *session*)
conceito (*concept*)	noção, conteúdo, regime, âmbito (*notion*, *content*, *regime*, *range*)	directiva, ingresso, nomeação, européia (*directive*, *ingress*, *appointment*, *european*)
categoria (*category*)	cargo, lugar, carreira, escalão (*post*, *position*, *career*, *grade*)	vencimento, nomeação, promoção, escalão (*expiration*, *appointment*, *promotion*, *grade*)
actividade (*activity*)	função, serviço, acção, cargo (*function*, *service*, *action*, *post*)	exercício, acção, serviço, processo (*exercise*, *action*, *service*, *process*)

Table 7. Similarity lists produced by two sorts of contexts: contexts with +prep and ↑↓–attributes and contexts with −prep and ↑–attributes (like Grefenstette's strategy).