Mapping Syntactic Dependencies onto Semantic Relations

Pablo Gamallo[1] **Marco Gonzalez**[2] **Alexandre Agustini**[3] **Gabriel Lopes**[4] and **Vera S. de Lima**[5]

Abstract. This paper presents a corpus-based method for extracting semantic relations between words. The method is based on two sequential procedures. First, it automatically classifies syntactic dependencies according to their selection restrictions. Those dependencies that require the same selection restrictions are put together into the same semantic group. Then, interpretation rules are applied on the classified syntactic dependencies, in order to learn the specific semantic relations underlying syntactically related words.

1 Introduction

The main aim of this paper is to describe a corpus-based approach to automatically extract semantic relationships between words. Our method is based on the assumption that syntactic dependencies contain conceptual regularities underlying semantic relations. This assumption is motivated by the observation that syntactic dependencies provide information on the semantic relations between the concepts (or meanings) the related words denote [18].

The approach is characterised by the following properties. On the one hand, it is based on a knowledge-poor learning strategy [9], since pre-encoded semantic information from external lexical resources (e.g., machine-readable dictionaries, handcrafted thesauri, . . .) is not required. In particular, our method relies on an unsupervised strategy for clustering semantically similar syntactic dependencies. This allows us to identify and put together those syntactic dependencies that presuppose similar semantic relationships. Such a method is close to some aspects of the clustering strategy used in the ASIUM system [4].

On the other hand, we do not need to previously define a set of lexico-syntactic patterns to extract semantic relations between words. In [10], patterns like $NP\, such\, as\, \{NP, ...\}$, $NP\, specially\, \{NP, ...\}$, and so on, are used to identify relations of hyperonymy. Work by [2] is able to identify meronymic relations using $NP\, of\, the\, NP$ and $NP\, in\, a\, NP$ patterns. Such methods may have good performance in a particular domain, but the cost of adapting the patterns to a new domain can be too hard [15]. Moreover, the scope of these patterns is rather limited. Instead of lexico-syntactic patterns, our method provides a set of symbolic interpretation rules. These rules are able to map syntactic dependencies onto the semantic relations that will be used to thesaurus design. Some of the ideas underlying the mapping rules are not far from some aspects of the approaches to semantic interpretation introduced in [18, 1].

[1] Dept. de Informática, Universidade Nova de Lisboa, Portugal.
[2] Faculdade de Informática, PUCRS, Brazil.
[3] Dept. de Informática, Universidade Nova de Lisboa, Portugal.
[4] Dept. de Informática, Universidade Nova de Lisboa, Portugal.
[5] Faculdade de Informática, PUCRS, Brazil.

This paper is organised as follows. We start by describing in section 2 the internal structure of syntactic dependencies, how they are identified from large corpora, as well as how they are clustered in semantic groups. In particular, we put together those dependencies sharing the same selection restrictions. Then, section 3 describes the interpretation rules used to map syntactic dependencies onto semantic roles. The semantic roles we used as output of the mapping rules are not limited to the four qualia roles of the Generative Lexicon Theory [16]. Then, section 4 outlines the main organising elements we need to design a particular relational thesaurus. Special attention will be paid to the representation of polysemous words, as well as to the comparison between our organising principles and the basic structure of WordNet. Finally, in section 5, we make a brief introduction of the application tasks (Information Retrieval and Robust Parsing) on which our method is being evaluated. A more accurate description of these application tasks is beyond the scope of the article.

The experiments presented in this paper were performed on 1,5 million of words belonging to the *P.G.R.* (*Portuguese General Attorney Opinions*) corpus, which is a domain-specific Portuguese corpus containing case-law texts. The fact of using a domain-specific corpus makes easier the learning task, given that we have to deal with a limited vocabulary with reduced polysemia.

2 Extracting Semantically Similar Syntactic Dependencies

We attempt to identify and describe the semantic relations embedded in syntactic dependencies. For this purpose, we start by extracting syntactic dependent words from partially analysed text corpora. In order to identify syntactic dependencies between words, we apply a two-step process. First, we select candidate dependencies by taking into account basic syntactic rules. Second, we build clusters of semantically similar dependencies. We assume that those dependencies that have been clustered represent true syntactic relations. These two different procedures, both selection of candidate syntactic dependencies and clustering of similar candidates, will be described in the following subsections.

2.1 Heuristics to Select Candidate Dependencies

The procedure for selecting candidate binary dependencies takes into account only morphologic and syntactic information. First, the corpus is tagged by the part-of-speech tagger presented in [13]. Then, the tagged corpus is analysed in sequences of basic chunks by using some of the potentialities of the shallow parser presented in [17]. The parser produces a single partial syntactic description of sentences, which are analysed as sequences of not related basic chunks (NP, PP, VP, . . .). Finally, we use simple heuristics based on right association in order to attach basic chunks: a chunk tends to be attached to another chunk immediately to its right. We consider that the word heads of two attached chunks form a candidate syntactic dependency. It can be easily seen that syntactic errors may appear since attachment heuristics do not take into account long distant dependencies.[6] Given that these heuristics rely on poor-defined linguistic rules, the correctness of attachments is not guarateed. For reasons of attachment errors, it is argued here that the binary dependencies identified by our elementary heuristics are mere hypotheses on attachment; hence they are mere *candidates*.

Let's describe the internal structure of a candidate 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 relation [8]: $(r; w1^{\downarrow}, w2^{\uparrow})$, where

- r can be instantiated by specific grammatical markers such as particular prepositions, subject relations, direct object relations, etc.;
- arrows"$\downarrow$" and "$\uparrow$" represent the *Head* and *Complement* position, respectively;
- $w1$ is the word in the Head position and $w2$ the word in the Complement position.

Binary dependencies denote grammatical relationships between the Head and its Complement. They are asymmetric relationships in such a way that only the syntactic properties of the Head are inherited by the composite construction. Nevertheless, in semantic terms, both words -the Head and the Complement- seem to be interdependent: The Complement must fill the semantic preferences (i.e., selections restrictions) required by the Head, and the Head also must fill the semantic preferences (i.e., selection restrictions) required by the Complement. We argue that a candidate dependency represents a true syntactic relation if, at least, one of these semantic requirements is identified and verified. The following subsection describes how we check whether two candidate related words are semantically interdependent.

2.2 Building Clusters of True Dependencies

We start by describing the bi-directional semantic requirements between the two syntactic positions constituting a dependency. Then, we describe how syntactic positions can be compared and then clustered forming semantic groups. Finally, we explain how this information is used to validate candidate dependencies. More precisely, a candidate dependency is considered a true relation if, at least, one of the two internal syntactic positions is involved in semantic clustering. This must be used as evidence to identify syntactico-semantic relationships.

2.2.1 Co-Restriction

In most linguistic research, a syntactic dependency between two words is semantically described as a simple restriction. The Head is viewed as the word that imposes semantic constraints (selection restrictions) on the Complement, which must fill such constraints. While the Complement is syntactically and semantically dependent, the Head keeps both a high degree of syntactic and semantic autonomy. Nevertheless, recent linguistic research assumes that the two words related by a syntactic dependency are mutually restricted by semantic constraints [16, 6]. We call the process of mutual restriction between two related words "co-restriction". Co-restriction is based on the fact that each word in a syntactic dependency both imposes semantic restrictions and matches semantic requirements. For instance, consider the expression `fase da evolução` (*phase of the evolution*), which occurs in corpus *P.G.R.* We assume that the candidate syntactic dependence $(de; fase^{\downarrow}, evolução^{\uparrow})$ extracted from that expression is semantically well-formed if, at least, one of the following two conditions are verified: either the word `evolução` (*evolution*) fills the selections restrictions imposed by the Head, or the word `fase` (*phase*) fills the selection restrictions imposed by the Complement.

These conditions address the problem of identifying selection restrictions. For this purpose, we aim at identifying both syntactic positions requiring similar selection restrictions, and those words appearing in similar syntactic positions. In the following, we introduce the main assumptions of this strategy.

2.2.2 Similar Syntactic Positions

Consider again the expression `fase da evolução` (*phase of the evolution*). Let's note $[\lambda x^{\uparrow}(de; fase^{\downarrow}, x^{\uparrow})]$ the syntactic position filled by the words appearing as Complements of `fase`, within the grammatical dependency introduced by preposition `de`. And let's note $[\lambda x^{\downarrow}(de; x^{\downarrow}, evolução^{\uparrow})]$ the syntactic position filled by the words appearing as Heads of `evolução`, within the dependency introduced by the same preposition. In corpus *P.G.R*, the words appearing in the position $[\lambda x^{\uparrow}(de; fase^{\downarrow}, x^{\uparrow})]$ refer to entities describing some kind of process or action, such as: `execução` (*execution*), `investigação` (*investigation*), `trabalho` (*work*), etc. On the other hand, the words appearing in $[\lambda x^{\downarrow}(de; x^{\downarrow}, evolução^{\uparrow})]$ refer to temporal or modal parts of processes: `momento` (*moment*), `período` (*period*), `resultado` (*result*), `fim` (*end*), etc. As has been said, not only the Head seems to semantically restrict the type of its Complement, but also the Complement somehow restricts the semantic type of the Head.

Following ideas presented in ([5, 4]), we assume that two different syntactic positions are semantically similar and, then, impose the same selection restrictions, if they require similar sets of words, i.e., if they have the same word distribution. Our algorithm uses a weighted Jaccard coefficient to measure similarity between word sets (see [7] for details). Note that we do not measure similarity between single words by comparing their syntactic distribution ([9, 12]. Rather, we measure similarity between syntactic positions by comparing their word distribution. In our corpus, the distribution of words appearing in the syntactic position $[\lambda x^{\uparrow}(de; fase^{\downarrow}, x^{\uparrow})]$ is similar to the distribution of words appearing in: $[\lambda x^{\downarrow}(em; x^{\downarrow}, curso^{\uparrow})]$, $[\lambda x^{\downarrow}(de; período^{\downarrow}, x^{\uparrow})]$, $[\lambda x^{\downarrow}(de; resultado^{\downarrow}, x^{\uparrow})]$, corresponding to the temporal expressions `em curso` (*in course*), `período de` (*period of*), `resultado de` (*result of*). Indeed, these syntactic positions seem to be semantically homogeneous. Likewise, the distribution of words appearing in $[\lambda x^{\downarrow}(de; x^{\downarrow}, evolução^{\uparrow})]$ is similar to the distribution of words appearing in: $[\lambda x^{\downarrow}(de; x^{\downarrow}, execução^{\uparrow})]$, $[\lambda x^{\downarrow}(de; x^{\downarrow}, investigação^{\uparrow})]$, $[\lambda x^{\downarrow}(de; x^{\downarrow}, trabalho^{\uparrow})]$, which correspond to the following event expressions: `da execução` (*of the execution*), `da investigação` (*da investigation*), `do trabalho` (*of the work*), etc. These syntactic positions turn out to be semantically homogenous.

The clustering algorithm presented in [7] put together both sim-

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

ilar syntactic positions and the words associated with them. Table 1 shows two different clusters: Cluster_1, which is constituted by both the positions similar to $[\lambda x^{\uparrow}(de; fase^{\downarrow}, x^{\uparrow})]$, and the words having the higher distribution through these positions; Cluster_2, which is constituted by both the positions similar to $[\lambda x^{\downarrow}(de; x^{\downarrow}, evolução^{\uparrow})]$, and the most significant words distributed along these positions.

Table 1. Illustration of two clusters of similar positions and related words

Cluster_1:	$\{[\lambda x^{\uparrow}(de; fase^{\downarrow}, x^{\uparrow})]\, [\lambda x^{\downarrow}(em; x^{\downarrow}, curso^{\uparrow})]$ $[\lambda x^{\downarrow}(de; período^{\downarrow}, x^{\uparrow})]\, [\lambda x^{\uparrow}(de; resultado^{\downarrow}, x^{\uparrow})]\} =$ {processo execução investigação trabalho} (*process execution investigation work*)
Cluster_2:	$\{[\lambda x^{\downarrow}(de; x^{\downarrow}, evolução^{\uparrow})]\, [\lambda x^{\downarrow}(de; x^{\downarrow}, execução^{\uparrow})]$ $[\lambda x^{\downarrow}(de; x^{\downarrow}, investigação^{\uparrow})]\, [\lambda x^{\downarrow}(de; x^{\downarrow}, trabalho^{\uparrow})]\} =$ {fase momento período resultado fim} (*phase moment period result fim*)

Since the words distributed along similar syntactic positions represent the extensional description of their semantic restrictions, they tend to become homogenous semantic classes. For instance, the Complement words {processo, execução , investigação, trabalho, ...} represent the extensional description of the semantic class required by the Head fase, within the grammatical relation introduced by de. Thus, they are good candidates to become co-hyponyms.

2.2.3 Checking Candidate Dependencies

Information on bi-directional semantic requirements turns out to be very useful for checking the validity of the candidate dependencies extracted by syntactic heuristics. For example, we decide whether the candidate $(de; fase^{\downarrow}, evolução^{\uparrow})$ is a correct syntactic attachment only if, at least, one of the two following conditions is verified: either evolução belongs to Cluster_1, or fase belongs to Cluster_2.

Table 2. True dependencies inferred from Cluster_1

$(de; fase^{\downarrow}, processo^{\uparrow})\ (em; processo^{\downarrow}, curso^{\uparrow})$ $(de; período^{\downarrow}, processo^{\uparrow})\ (de; resultado^{\downarrow}, processo^{\uparrow})$ (*phase of the process, process in course, periode of the process, result of the process*)
$(de; fase^{\downarrow}, execução^{\uparrow})\ (em; execução^{\downarrow}, curso^{\uparrow})$ $(de; período^{\downarrow}, execução^{\uparrow})\ (de; resultado^{\downarrow}, execução^{\uparrow})$ (*phase of the execution, execution in course, periode of the execution, result of the execution*)
$(de; fase^{\downarrow}, investigação^{\uparrow})\ (em; investigação^{\downarrow}, curso^{\uparrow})$ $(de; período^{\downarrow}, investigação^{\downarrow})(de; resultado^{\downarrow}, investigação^{\uparrow})$ (*phase of the investigation, investigation in course, periode of the investigation, result of the investigation*)
$(de; fase^{\downarrow}, trabalho^{\uparrow})\ (em; trabalho^{\downarrow}, curso^{\uparrow})$ $(de; período^{\downarrow}, trabalho^{\downarrow})\ (de; resultado^{\downarrow}, trabalho^{\uparrow})$ (*phase of the work, work in course, periode of the work, result of the work*)

That means that two words are considered as syntactically dependent if, at least, one of them may be semantically required by the other. As word fase is a word member of Cluster_2, we may infer that it is semantically required by evolução. Both words are, then, syntactically and semantically related. Indeed, a single semantic restriction may be used as a significant evidence for establishing that two words are syntactically dependent. On this basis, we can use the semantic restriction underlying Cluster_1, illustrated in Table 1, as evidence for considering all dependencies of Table 2 as true syntactic relationships.

In summary, the process of checking whether a candidate dependency is a true syntactic attachment leads us to put together dependencies having the same semantic requirements and then describing the same conceptual relations. As syntactic dependencies are organised in semantically homogenous clusters, we may map them to semantic relations in a more systematic way. In the next section, we describe the rules and filters we use to extract semantic information such as conceptual relations and semantic roles, from syntactic dependencies.

3 Mapping from Syntactic Dependencies to Semantic Relations

3.1 Meaning of Grammatical Relators

In the previous section, we explained how semantic regularities underlying syntactically related words were used to calculate similarity between syntactic dependencies. In the following, our aim is to describe a method for extracting the specific semantic information carried by syntactic dependencies.

Research on cognitive linguistics (e.g., [11]) is laying great stress to the semantic interpretation of syntactic relators (e.g., grammatical functions, prepositions, ...). Two opposite approaches to the semantics of syntactic relators are often reported. On the one hand, it is argued that syntactic relators are endowed with very schematic senses that are elaborated in context. On the other hand, it is claimed that they are polysemic, i.e., they have several related meanings. In both cases, however, the most relevant linguistic task is not to describe either the schematic sense or the related meanings of a syntactic relator, but either how the schematic sense is elaborated or how a specific meaning is selected. In particular, many works focus on enumerating and organising the linguistic information required to correctly interpret the specific meaning of syntactically related words. On this basis, we do not attempt to define the meaning of syntactic relators. Rather, we aim at defining how a syntactic relator cooperates with other available linguistic information (e.g., morphosyntactic category of the related words, their lexical content, co-textual data), in order to interpret the semantic relation underlying two syntactically dependent words.

3.2 Linguistic Constraints and Interpretation Rules

Interpretation rules map syntactic dependencies onto semantic representations. In particular, they are able to extract semantic information by taking into account different internal aspects of dependencies, such as for instance syntactic relators, morphosyntactic markers, etc. These internal aspects are viewed by the interpretation rules as constraints filtering semantic representations. Separately, none of the constraints is able to filter a single meaning, but when they are put together, their filtering activity becomes more efficient. For example, preposition de (*of/by*) is hardly associated to some kind of semantic information since it is too much ambiguous. However, when further information on the Head and the Complement related by the preposition is available, a compositional semantic interpretation becomes possible.

The degree of specification of the semantic information associated to syntactic dependencies relies on the type of linguistic constraints used by the interpretation rules. Since we do not dispose of lexical resources, we only consider the following three types of constraints: (1)

syntactic relators (subject, direct object, and prepositions); (2) morphosyntactic categories of the two related words (verb and noun); /3) presence or absence of determinant in the Complement. These constraints are put together in order to build "grammatical patterns" of syntactic dependencies. For instance, we use the notation $de+$ for representing the pattern constituted by: (1) preposition `de`; (2) presence of the determiner before the Complement; (3) the Head and the Complement are both nouns. If the determiner is not present before the Complement, we use the notation $de-$. When the symbols "+" and "-" are removed from the pattern, the determiner can be either present or absent. In case the Head is a verb and the Complement a noun, we note $iobj\_de$. There are also patterns without preposition, namely $subj$ and $dobj$, which describe respectively the relation of subject and direct object between a verb (the Head) and a noun (the Complement).[7]

Let's take an example. Consider the expressions `resultado da nomeação`, (*result of the nomination*), `nomeação do presidente` (*nomination of the president*), and `gabinete do presidente` (*president's office*) . At the grammatical level, the three expressions contain pattern $de+$. In Portuguese, the semantic interpretation of this pattern seems to be close to the very abstract relation of possession, more precisely, the Complement must be viewed as the "possessor" of the Head. Hence, we may infer that `nomeação` is the possessor of `resultado` (i.e., nominations have results), as well as `presidente` is the possessor of `nomeação` and `gabinete` (i.e., presidents have nominations and offices). Note that if a particular constraint is changed, then the meaning of the pattern may vary in a significant way. For instance, if the determiner of the Complement is removed (i.e., if we have the pattern $de-$), we can obtain a larger range of semantic contents: namely, the reverse relation of possession (`expositor de vidro` - *stand made of glass*), the relation of hyperonymy (`processo de nomeação` - *process of nomination*), or even the notion of modality (`dimensão de vulto` - *important dimension*). Likewise, if the Head is not a noun but a verb (i.e., if we have the pattern $iobj\_de$), the Complement appears as being an external participant of the event denoted by the verb. Such a participant, called here "external theme", is neither the agent nor the affected patient (i.e., internal theme) of the event. Rather, it may embrace the notions of transferred object, instrument, coadjuvant, goal, etc.

Further information, such as specific lexical content of the related words, co-textual data, etc., could also be used as constraints to elaborate in a more accurate way the meaning of dependent expressions. In the following, interpretation rules will be merely defined on the basis of grammatical constraints.

3.3 The Semantic Space of the Interpretation Rules

The semantic information associated to grammatical patterns is distributed along a scale of gradient relational concepts, such as hyperonymy, possession, location, modality, causality, agentivity, etc. Since there is no means for clearly dividing them into discrete components, the semantic space should be essentially viewed as a continuum of fuzzy notions [11]. We must stress the importance of two properties of this space. First, interpretation rules do not map grammatical patterns onto specific values of the space, but on a range of neighbouring values organised around a prototypical core. Second, the level of specification of the scalar values is far from being rigid, it can be changed and adapted according to the particular application it will be used for.

Let's see Figure 1, which depicts two vertical axes and two horizontal axes. Each axis represents a scale of semantic roles. On the one hand, the vertical axes (noted **C**) contain the semantic roles that Complements can play in various grammatical patterns. The vertical axis on the left represents the scale of roles (i.e., Hyperonymy, Possessed, Possessor, and so on) that we propose for interpreting the Complements of syntactic dependencies. The scale on the right is organised by the qualia roles (i.e., Formal, Constitutive, Telic, Agentive) which are described in the Generative Lexicon Theory [16]. Note that a great part of this scale of values remains underspecified. Indeed, only "Constitutive" seems to be adequate to characterise the possible semantic roles that Complements can play in the grammatical patterns. This leads us to infer that the semantic space organised by the qualia roles is not specific enough to be used in the process of interpreting syntactic dependencies.

On the other hand, Figure 1 depicts two horizontal axes (noted **H**), containing the semantic roles that the Heads can play in the grammatical patterns. The axis situated at the bottom contains the semantic roles that we propose to interpret the Heads of syntactic dependencies, that is, Hyperonym, Possessor, Possessed, Located, Effect, Purpose, etc. They are complementary to the values appearing in the scale of Complements (the vertical axis on the left). The axis at the top is organised by the qualia roles. Even though it is more elaborate that the scale of Complements, it still contains several underspecified zones.

Figure 1 can be conceived as the visual counterpart of the interpretation rules we define for some Portuguese grammatical patterns. In this figure, grammatical patterns are represented as squares occupying some regions within the semantic space. The larger the size of the square, the more abstract the semantic content associated to the pattern. Consider, for example, the spatial region representing the semantic values of pattern $de-$. This region embraces a large range of values: it goes from Hyponym to Modality, even if the roles extracted from the relation of possession should remain more typical. Concerning the scales organised by the qualia roles, it occupies the large region attributed to the Formal and Constitutive roles. Pattern $subj$, which describes the relation of subject between a verb (the Head) and a noun (the Complement), also embraces a large amount of values: it occupies the space from Cause to Internal Theme in the scale of Complements. By contrast, some of the patterns may be associated to small regions of values. For instance, patterns $a+$, $de+$, and $para$ are mapped both onto the specific role "Possessor" in the scale of Complements and onto the role "Possessed" in the scale of Heads. Likewise, pattern $dobj$ merely occupies the space associated to the role of Theme (external and internal) in the scale of Complements. Note also that the same semantic region can be occupied by a great number of grammatical patterns, for instance, $iobj\_a-$, $a-$, $iobj\_em$, em, $iobj\_sobre$, $sobre$. All of them are mapped onto the same region: the large space embracing the relation of location, which encloses in the boundaries some aspects of possession and modality.

The syntactico-semantic information displayed by Figure 1 is organised in a symbolic way by means of interpretation rules such as:

$$x{=}Possessed\,;\,y{=}Possessor \quad \Longrightarrow$$
$$\left[\lambda x^{\downarrow}\lambda y^{\uparrow}(de{+};x^{\downarrow},y^{\uparrow})\right] or \left[\lambda x^{\downarrow}\lambda y^{\uparrow}(a{+};x^{\downarrow},y^{\uparrow})\right] or \left[\lambda x^{\downarrow}\lambda y^{\uparrow}(para;x^{\downarrow},y^{\uparrow})\right]$$

By means of this rule, the Heads (expressed by the variable *x*) of the patterns $de+$, $a+$, and $para$ are mapped onto the semantic role

[7] Note that we only use patterns constraining the Complement to be a noun. Verbal and clausal complements were not taken into account by the attachment heuristics described above because their syntactic behaviour is not trivial.

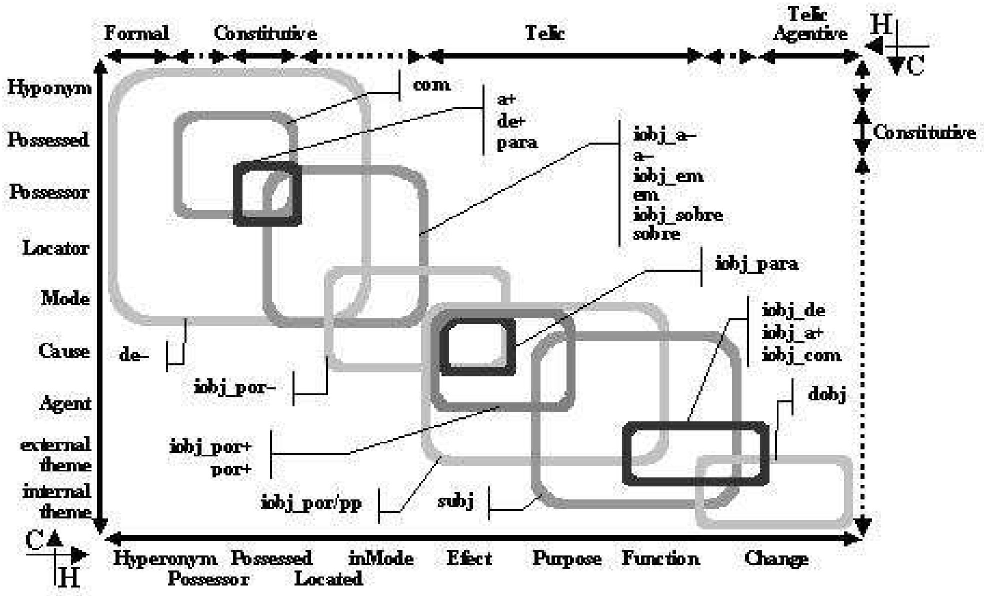

Figure 1. A proposal of a semantic space for portuguese grammatical patterns

"Possessed", whereas the Complements (expressed by *y*) are mapped onto the role "Possession".

The schematic semantic content obtained by this kind of rules (e.g. possession, location, etc.), can be viewed as filters driving the interpretation of complex expressions. Specific and elaborate interpretations may be achieved only if we have access to more precise constraints, such as lexico-semantic information of the Head and the Complement, as well as further contextual information. However, unlike work by [18, 1], we do not use such rules to interpret dependent expressions, but to describe general principles organising thesaurical relations between words. In the following section, we propose an automatic method to build a thesaurus from corpora, by organising in a principled way the semantic relations between words.

4 Thesaurus Design

4.1 Classets and Semantic Relations

The structure of our thesaurus is analogous, but not identical, to the relational organisation proposed by WordNet [14]. In WordNet, the basic semantic relation between word forms is synonymy. Two words are considered as synonyms if they have the same meaning in a particular context. Synonyms are gathered in *synsets*, which are sets of words linked by a common contextual meaning. A synset do not explain what a meaning is, it merely point up a particular meaning. A word is polysemic (i.e., has various contextual meanings), if it belongs to different synsets. In WordNet, there is only one semantic relation between word forms: synonymy. The other relations, e.g., hyperonymy, meronymy, antonymy, etc., are relations between contextual meanings, i.e., between synsets.

In our framework, words are not clustered into groups of contextual synonyms (i.e., in synsets), but into groups constituted by words filling the selection restrictions of semantically similar syntactic positions. The words clustered in this way share the same combinatorial properties, and then have similar syntactic behaviour. Section 2.2 showed how we generate these sets of words. In analogy to the notion of synset, we call them *classets*. Given that they are grouped according to the semantic preferences of syntactic positions, classets are not only formed by synonyms. For instance, in the classet {`processo, procedimento, execução, trabalho, ...`}, words are not related by synonymy but by class membership. This way, a classet represents the common semantic class of the classet members, which are co-hyponyms. The classet may be viewed as a particular contextual word meaning. Note that co-hyponymy is a more general relation than synonymy: even if all synonyms are co-hyponyms, the reverse is not true.

Like synsets, classets are used to characterise the semantic relations between words. Semantic relations are defined as relations between classet members and their syntactically dependent words. The interpretation rules defined above allow us to precise what semantic relations (and semantic roles) underlie the syntactic dependencies between classet members and their related words. Note that in our approach, crosscategorial semantic relations are allowed. For example, classets of nouns are related, not only to nouns, but also to verbs by means of semantic theta-roles such as "agent", "location", or "theme". In [3], it is also described a method to extracting crosscategorial semantic relations between nouns and verbs. This kind of semantic information is unfortunately absent of WordNet, because only intracategorial semantic relations are allowed.

Table 3 shows at the top the larger classet built by our clustering algorithm. Its members are co-hyponyms of the general class referring to legal documents. In addition, this table also displays three types of information associated with the classet: (1) the words syntactically dependent on the classet members (column on the left); (2) the syntactic positions used to relate the classet members to their dependent words (middle column); (3) the semantic roles assigned by our interpretation rules to the classet members regarding their associated syntactic positions (column on the right). In order to simplify the description, the grammatical patterns of the syntactic positions do not contain information concerning the presence or absence of determiners before the Complement. The roles on the right are the more prototypical values obtained by applying the interpretation rules on these grammatical patterns. On this basis, the classet constituted by legal documents such as *articles, decrees*, or *laws* are the Possessors of objects like *annex, chapter, letter, text, content*, or *force*. They can also play the role of Theme in actions like *finish, add, modify, edit*, or *ratify*; they are the External Theme of *emanate_from*, *be_in_conflict_with*; the Locator of *enumerate_in, mention_in, state_in*; and the Agent of *fix, introduce, report*, or *prescribe*.

4.2 Representation of Polysemic Words

Following the structure of WordNet, a word form is considered as polysemic if it belongs to various classets. The classets to which a polysemic word belongs represent its different, even if related, meanings. Consider the word `trabalho` (*work/job*). Our clustering strategy integrates this word into various word clusters, i.e., classets. Tables 4, 5, and 6 show information concerning three of these classets. The classet illustrated in Table 4 aggregates words referring to temporal actions. Indeed, the co-hyponyms constituting the classet appears to be syntactically and semantically associated with the words `fase` (*phase*), `suspensão` (*interruption*), and `curso` (*course*) in the binary dependencies `fase do trabalho` (*phase of the work*), `suspensão do trabalho` (*interruption of the work*), and `trabalho em curso` (*work in course*). In semantic terms, `trabalho` is viewed here as an action related to words referring to its temporal internal facets. In Table 5 the classet seems to describe, not temporal actions, but the result of an action. Such a meaning becomes salient in the syntactic position $[\lambda x^{\uparrow}(iobj\_por; receber^{\downarrow}, x^{\uparrow})]$ (*to receive in payment for*). Indeed, the cause of receiving money is not the action of working, but the object done or the state achieved by working. Finally, Table 6 illustrates the more typical meaning of `trabalho`: it is a job, function or task, which can be carried out by professionals. This is why the cohyponyms of this classet are syntactically and semantically related to words like `inspector` (*inspector*) and `reviser` (*supervisor*).

The three contextual meanings of `trabalho` represent three related aspects of the concept of working. Yet, only one of them seems to be relevant in a particular compositional use of the word. Our method relies on the fact that each syntactic position in which a word appears is able to select one of its contextual meanings.

5 Evaluation and Applications

The results of our semantic acquisition method are being used for two different tasks. On the one hand, the selection restrictions acquired by our clustering strategy are introduced in the lexicon as semantic patterns of subcategorisation. This subcategorisation information is used to correct the spurious syntactic attachments proposed by the parser at the first analysis step. Once the subcategorisation information is learned, we apply a diagnoser parser on the candidate dependencies to check whether they are correct or not. For the odd attachments, the diagnoser proposes new hypotheses provided that it is endowed with the appropriate subcategorisation information. On the other hand, our results are currently being integrated to an IR system for precise and rapid location of documents in the *P.G.R.* corpus collection (*http://coluna.di.fct.unl.pt/∼pgrd*). The acquired word semantic relations are used to extend and improve documents recall. The degree of efficacy in these applications may serve as a reliable evaluation for measuring the soundness of our learning strategy.

6 Summary

This paper has presented a corpus-based method for extracting semantic relations between words. This method is based on two sequential procedures. First, it automatically classifies syntactic positions according to their selection restrictions. Those positions that impose the same selection restrictions are put together into the same semantic group. Furthermore, we also identify groups of nouns according to their distribution through syntactic positions with the same selection restrictions. We called *classets* such groups of nouns. Then, once classets and classes of similar syntactic positions have been learned, interpretation rules are applied on them in order to learn the specific semantic relations underlying syntactically related words. Finally, these semantic relations are organised in a thesaurical structure, which has significant analogies to the WordNet structure.

ACKNOWLEDGEMENTS

This work is supported in part by grants of FCT - PRAXIS XXI, Portugal; CAPES, Brazil; PUCRS, Brazil; and the MLIS 4005 European project TRADAUT-PT.

REFERENCES

[1] Carol A. Bean, Thomas C. Rindflesch, and Charles A. Sneiderman, ‘Automatic semantic interpretation of anatomic spatial relationships in clinical text’, in *AMIA Annual Fall Symposium*, pp. 897–901, (1998).

[2] M. Berland and E. Charniak, ‘Finding parts in very large corpora’, in *ACL’99*, (1999).

[3] P. Bouillon, C. Fabre, P. Sébillot, and C. Jacquemin, ‘Apprentissage de ressources lexicales pour l’extension de requêtes’, in *T.A.L. pour les Recherches d’Information*, pp. 367–393. Hermès, (2000).

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

[5] David Faure and Claire Nédellec, ‘Asium: Learning subcategorization frames and restrictions of selection’, in *ECML’98*, (1998).

[6] Pablo Gamallo, *Construction conceptuelle d’expressions complexes: traitement de la combinaison nom-adjectif*, Ph.D. dissertation, Université Blaise Pascal, Clermont-Ferrand, France, 1998.

[7] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes, ‘Selection restrictions acquisition from corpora’, in *10th Portuguese Conference on Artificial Intelligence (EPIA’01)*, pp. 30–43, Porto, Portugal, (2001). LNAI, Springer-Verlag.

[8] Pablo Gamallo, Caroline Gasperin, Alexandre Agustini, and Gabriel P. Lopes, ‘Syntactic-based methods for measuring word similarity’, in *Text, Speech, and Discourse (TSD-2001)*, eds., V. Mautner, R. Moucek, and K. Moucek, 116–125, Berlin:Springer Verlag, (2001).

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

[10] Marti A. Hearst, ‘Automatic acquisition of hyponyms from large text corpora’, in *COLING’92*, pp. 539–545, Nancy, (1992).

[11] Ronald W. Langacker, *Foundations of Cognitive Grammar: Descriptive Applications*, volume 2, Stanford University Press, Stanford, 1991.

[12] Dekang Lin, ‘Automatic retrieval and clustering of similar words’, in *COLING-ACL’98*, Montreal, (1998).

[13] Nuno Marques, *Uma Metodologia para a Modelação Estaística da Subcategorização Verbal*, Ph.D. dissertation, Universidade Nova de Lisboa, Lisboa, Portugal, 2000.

[14] G. Miller, ‘Wordnet: an on-line lexical database’, *International Journal of Lexicography, 4(3)*, (1990).

[15] Emmanuel Morin and C. Jacquemin, ‘Projecting corpus-based semantic links on a thesaurus’, in *ACL’99*, pp. 20–26, Meryland, USA, (1999).

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

[17] V. Rocio, E. de la Clergerie, and J.G.P. Lopes, ‘Tabulation for multipurpose partial parsing’, *Journal of Grammars*, **4**(1), (2001).

[18] Martin Romacker, Katjia Markert, and Udo Hahn, ‘Lean semantic interpretation’, in *IJCAI’99*, pp. 868–875, Stockold, Sweden, (1999).

Table 3. Classet of legal documents

alínea artigo código constituição convenção decreto diploma disposição estatuto legislação lei norma preceito regulamento *(paragraph article code constitution aggreement decree diploma disposition statute legislation law norm precept regulation)*		
Related Words (RW)	Syntactic Positions	Semantic Roles
anexo infracção *(annex infraction)*	$[\lambda x^{\uparrow}(a;RW^{\downarrow},x^{\uparrow})]$	$x = Possessor$
força *(force)*	$[\lambda x^{\downarrow}(com;x^{\downarrow},RW^{\uparrow})]$	$x = Possessor$
alcance capítulo contexto emisão escopo letra redacção teor texto *(scope chapter context emission scope letter redaction content text)*	$[\lambda x^{\uparrow}(de;RW^{\downarrow},x^{\uparrow})]$	$x = Possessor$
acabar aditar alterar editar fazer ratificar reger regulamentar violar *(finish add change modify edit make ratify rule regularise violate)*	$[\lambda x^{\uparrow}(dobj;RW^{\downarrow},x^{\uparrow})]$	$x = Theme$
colidir conjugar *(be-in-conflict conjugate)*	$[\lambda x^{\uparrow}(iobj\_com;RW^{\downarrow},x^{\uparrow})]$	$x = External\ Theme$
emanar *(emanate)*	$[\lambda x^{\uparrow}(iobj\_de;RW^{\downarrow},x^{\uparrow})]$	$x = External\ Theme$
afirmar conter enumerar enunciar estabelecer estipular indicar mencionar *(assert contain enumerate state stablish indicate mention)*	$[\lambda x^{\uparrow}(dobj;RW^{\downarrow},x^{\uparrow})]$	$x = Locator$
adoptar alterar definir fixar introducir reger regular revogar *(adopt modify define appoint introduce rule regulate revoke)*	$[\lambda x^{\uparrow}(iobj\_por{+}pp;RW^{\downarrow},x^{\uparrow})]$	$x = Agent$
incompatibilidade matéria *(incompatibility matter)*	$[\lambda x^{\downarrow}(sobre;x^{\downarrow},RW^{\uparrow})]$	$x = Located$
admitir atribuir prescrever producir regular reportar resaltar respeitar alterar definir fixar introducir reger revogar *(admit ascribe prescribe produce regulate allude heighten modify define appoint introduce rule revoke)*	$[\lambda x^{\uparrow}(subj;RW^{\downarrow},x^{\uparrow})]$	$x = Agent$

Table 4. Classet of trabalho as a temporal entity

contrato execução exercício prazo procedimento processo trabalho *(agreement execution practice term/time procedure process work)*		
Related Words (RW)	Syntactic Positions	Semantic Roles
fase suspensão conclusão *(phase interruption end)*	$[\lambda x^{\uparrow}(de;RW^{\downarrow},x^{\uparrow})]$	$x = Possessor$
curso *(course)*	$[\lambda x^{\downarrow}(em;x^{\downarrow},RW^{\uparrow})]$	$x = Located$
controlar *(control/supervise)*	$[\lambda x^{\uparrow}(em;RW^{\downarrow},x^{\uparrow})]$	$x = Theme$

Table 5. Classet of trabalho as the result of an action

contrato exercício prestação recurso serviço trabalho *(agreement practice instalment appeal service work)*		
Related Words (RW)	Syntactic Positions	Semantic Roles
retribuição *(remuneration)*	$[\lambda x^{\uparrow}(de;RW^{\downarrow},x^{\uparrow})]$	$x = Possessor$
receber *(receive)*	$[\lambda x^{\uparrow}(iobj\_por;RW^{\downarrow},x^{\uparrow})]$	$x = Cause$
remuneração *(remuneration)*	$[\lambda x^{\uparrow}(por;RW^{\downarrow},x^{\uparrow})]$	$x = Cause$

Table 6. Classet of trabalho as a job

actividade atribuição cargo exercício função lugar trabalho *activity attribution post practice function post workjob)*		
Related Words (RW)	Syntactic Positions	Semantic Roles
desempenho incompatibilidade *(performance incompatibility)*	$[\lambda x^{\uparrow}(de;RW^{\downarrow},x^{\uparrow})]$	$x = Possessor$
revisor inspector *(reviser supervisor)*	$[\lambda x^{\downarrow}(de;x^{\downarrow},RW^{\uparrow})]$	$x = Possessed$
desempenhar remunerar *(accomplish remunerate)*	$[\lambda x^{\uparrow}(dobj;RW^{\downarrow},x^{\uparrow})]$	$x = Theme$
investir *(invest)*	$[\lambda x^{\uparrow}(iobj\_em;RW^{\downarrow},x^{\uparrow})]$	$x = Locator$
perceber *(receive-fees)*	$[\lambda x^{\uparrow}(subj;RW^{\downarrow},x^{\uparrow})]$	$x = Agent$