Extraction of Complex Index Terms in Non-English IR: A Shallow Parsing Based Approach

Jesús Vilares [a,*], Miguel A. Alonso [a], Manuel Vilares [b]

[a] *Department of Computer Science, University of A Coruña, Campus de Elviña, 15071 – A Coruña, Spain*

[b] *Department of Computer Science, University of Vigo, Campus As Lagoas s/n, 32004 – Ourense, Spain*

Abstract

The performance of Information Retrieval systems is limited by the linguistic variation present in natural language texts. Word-level Natural Language Processing techniques have been shown to be useful in reducing this variation. In this article, we summarize our work on the extension of these techniques for dealing with phrase-level variation in European languages, taking Spanish as a case in point. We propose the use of syntactic dependencies as complex index terms in an attempt to solve the problems deriving from both syntactic and morpho-syntactic variation and, in this way, to obtain more precise index terms. Such dependencies are obtained through a shallow parser based on cascades of finite-state transducers in order to reduce as far as possible the overhead due to this parsing process. The use of different sources of syntactic information, queries or documents, has been also studied, as has the restriction of the dependencies applied to those obtained from noun phrases. Our approaches have been tested using the CLEF corpus, obtaining consistent improvements with regard to classical word-level non-linguistic techniques. Results show, on the one hand, that syntactic information extracted from documents is more useful than that from queries. On the other hand, it has been demonstrated that by restricting dependencies to those corresponding to noun phrases, important reductions of storage and management costs can be achieved, albeit at the expense of a slight reduction in performance.

Key words: Information retrieval; Linguistic variation; Natural language processing; Shallow parsing; Finite-state transducers

[*] Corresponding author: tel. +34 981 167 000 ext. 1377, fax +34 981 167 160.
Email addresses: jvilares@udc.es (Jesús Vilares), alonso@udc.es (Miguel A. Alonso), vilares@uvigo.es (Manuel Vilares).

1. Introduction

Natural Language Processing (NLP) has frequently attracted the attention of the Information Retrieval (IR) community, since textual IR can be considered an NLP problem. This is because the task of deciding about the relevance of a given document with respect to a query basically consists in deciding whether the text of the document satisfies the information need expressed by the text of the query, the content of the document thus needing to be understood to a certain extent (Strzalkowski, 1999). This reasoning is supported by the fact that one of the major limitations of IR systems is the *linguistic variation* inherent to human language (Arampatzis et al., 2000; Galvez et al., 2006), i.e. the different linguistic alterations a term may suffer resulting in the non-possibility of matching, thereby reducing both precision and recall. Examples of these kinds of alteration include inflection, e.g. *mouse* vs. *mice*; the use of synonyms, e.g. *killer* vs. *murderer*; or the existence of syntactic transformations, e.g. *climatic changes* vs. *changes in the climate*. In recent years, progress in the field of NLP has resulted in the development of a new generation of efficient, robust and precise tools. These advances, together with the increasing power of new computers, facilitate the application of NLP systems in real IR environments.

Morphological variation, for example, has usually been solved in IR systems through the employment of stemmers, which reduce the word to its supposed grammatical root, or stem, through suffix stripping based on a list of frequent suffixes. The effectiveness of stemming is dependent on the morphological complexity of the language (Arampatzis et al., 2000). In the case of Romance languages, stemming does not seem to be an appropriate solution since many inflectional phenomena cannot be managed by such a simple tool. Moreover, stemming can also cause problems in further processing because of the loss of information it involves (Kowalski, 1997), a major problem when we intend to perform further processing.

In this context, NLP-based approaches seem to be a better solution. This is the case, for example, of the use of lemmatization as an alternative to stemming in order to eliminate inflectional variation (Vilares et al., 2004a): firstly, a linguistically motivated preprocessing (Graña et al., 2002) (tokenization, contraction splitting, separation of enclitic pronouns from verbal stems, proper noun recognition, etc.) is applied to avoid erroneous behaviors during further processing (Manning and Schütze, 1999; Jurafsky and Martin, 2000); secondly, a state-of-the-art part-of-speech tagger and lemmatizer is applied and the lemmas of the content words (Jacquemin and Tzoukermann, 1999), namely nouns, verbs and adjectives, the grammatical categories containing the main semantics of the text, are extracted for indexing, where the *lemma* of a word stands for its canonic form —e.g. infinitive in the case of verbs or masculine singular in the case of nouns and adjectives.

Going one step further, another possibility consists, for example, in using morphological families for dealing with derivational morphology (Vilares et al., 2001a). A *morphological family* is the set of words obtained from the same morphological root through derivation mechanisms (Jacquemin and Tzoukermann, 1999), taking into account the derivational morphemes, their allomorphic variants and the phonological conditions they must satisfy. For each family, automatically generated in an off-line process, a unique identifier or representative is selected. In this way, during indexing, such a representative will be used for all words belonging to the same family.

An important fact is that the computational complexity of this NLP processing chain is linear with respect to the length of the text, the increase in running cost with respect to stemming therefore being negligible (Vilares et al., 2001a).

At a higher level of linguistic description, we can find *syntactic variation*, the modification of the syntactic structure of a sentence whilst keeping its underlying semantics for retrieval purposes. Syntactic variation has been traditionally managed through two different approaches: the use of syntactic structures, and the use of phrases as complex index terms. The goal pursued is to increase the precision of retrieval, trying to avoid the limitations of the bag-of-terms paradigm (Tzoukermann et al., 1997). The use of complex representations based on syntactic structures such as trees (Smeaton et al., 1995; Sheridan and Smeaton, 1992; Vilares et al., 2001b, 2004b) or graphs (Montes-y-Gómez et al., 2000), for indexing and retrieval is not appropriate for large-scale use in real environments because of its high processing cost. A more feasible approach consists in the use of phrases as index terms, since phrases denote more meaningful concepts or entities than single words, thus being more precise and descriptive (Fagan, 1987; Strzalkowski and Perez-Carballo, 1994; Arampatzis et al., 2000) without damaging recall, because the simple terms compounding the phrase would also have matched (Mitra et al., 1997). Two types of phrase have traditionally been considered in IR: *statistical* phrases, obtained through statistical techniques looking for sequences of contiguous words co-occurring with a significant frequency (Mittendorfer and Winiwarter, 2001; Buckley et al., 1993; Fagan, 1987); and *syntactic* phrases, obtained through NLP techniques and formed by syntactically related sets of words (Narita and Ogawa, 2000; Koster, 2004; Jacquemin, 2001; Perez-Carballo and Strzalkowski, 2000; Hull et al., 1997; Smeaton et al., 1995; Sheridan and Smeaton, 1992; Fagan, 1987; Dillon and Gray, 1983), although most current syntactic solutions use only noun phrases as complex index terms (Kraaij and Pohlmann, 1998; Mitra et al., 1997; Hull et al., 1997; Fagan, 1987). Which phrase type has a better performance in IR tasks is still an undecided issue, although some results show syntactic phrases are the best choice when accurate parsing and syntactic disambiguation techniques are available (Arampatzis et al., 2000). Another common approach consists in limiting the complexity of complex terms by only using pairs, thus decompounding those complex terms formed by more than two single terms into compounds of only two terms (Koster, 2004; Arampatzis et al., 2000; Perez-Carballo and Strzalkowski, 2000; Fagan, 1987).

Finally, since the use of phrases alone as index terms enables only a partial view of the document to be captured (Mitra et al., 1997), complex index terms are frequently used in combination with simple index terms (Narita and Ogawa, 2000; Mitra et al., 1997; Hull et al., 1997; Strzalkowski and Perez-Carballo, 1994; Smeaton et al., 1995; Sheridan and Smeaton, 1992; Buckley et al., 1993; Fagan, 1987).

In this context, the aim of the present article is to study the viability of the application of NLP for dealing with syntactic variation in European IR systems, taking Spanish as a case in point. The greater linguistic complexity of Romance languages in comparison with English, at both a syntactic and morphological level, does not allow a direct extrapolation of the results so far obtained for English (Vilares et al., 2004a); these languages therefore demand their own experiments. Moreover, our approach introduces numerous features, which will be described in the following sections, that serve to differentiate our proposal from previous works.

Thus, in contrast to previous syntactic-based approaches, our proposal not only deals with strict syntactic variation, but also with morpho-syntactic variation, which combines both syntactic transformations and derivational phenomena. On the other hand, although previous approaches mainly work with noun phrases only, our work extends the range of phrases by also including verb phrases.

Another interesting aspect that differentiates our work from previous approaches is the fact that we have focused our study not only on the syntactic information obtained from queries, but also on the syntactic information offered by documents. Such syntactic information is obtained through shallow parsing, an approach often used in Information Extraction but not so commonly in IR.

Furthermore, we have to face one of the main problems in non-English NLP research, namely the lack of freely available linguistic resources: large tagged corpora, treebanks and advanced lexicons are not currently available for languages such as Spanish. The solution for minimizing this problem consists in restricting the complexity of the solutions proposed by focusing on the use of lexical information, which is easier to obtain. Limiting the complexity of the proposed approaches results in a general architecture which can be applied to Romance languages in particular and to languages with similar characteristics and behavior in general.[1] Furthermore, in order to minimize the computational cost of our approaches for their application in practical environments, finite-state technology has been widely used.

The structure of the rest of this article is as follows. Firstly, Section 2 explains how phrases can be used as complex index terms for dealing with syntactic and morpho-syntactic variation. Next, Section 3 describes the shallow parser developed for this task, while Section 4 introduces the indexing mechanism used. A detailed evaluation of our proposals is performed in Section 5. Finally, Section 6 explains our conclusions and future work.

2. Complex Terms as Index Terms

We can find noun and verb phrases in a text (Koster, 2004; Arampatzis et al., 2000). In both cases, the degree of specificity of the phrase is greater than those for the individual simple terms it contains, phrases therefore becoming more precise and descriptive index terms than their individual components (Fagan, 1987; Strzalkowski and Perez-Carballo, 1994; Arampatzis et al., 2000). Thus, phrases are often used to obtain *complex index terms*, also called *multi-word index terms*, in order to complement the semantics of documents captured by representations based on simple index terms. Accordingly, phrases to be used as complex index terms should include, at the very least, noun phrases together with those relations also involving verb phrases (i.e., involving the main verb of the sentence and its subject, object or adjuncts).

[1] Although this article focuses on Spanish, our approach has also been applied to Galician, a minority language that shares official status with Spanish in the region of Galicia, Northwest Spain. Portuguese and Galician developed separately as from the 14th century from a common language, Galaico-Portuguese, which was spoken in the ancient Kingdom of Galicia —now Galicia and Northern Portugal—, with Portuguese spreading southwards after unification of the country in the 14th century. It would therefore be a straightforward matter to also apply our approach to Portuguese.

Since the number of possible phrases is almost unlimited, the space of multi-word terms is much sparser than in the case of simple terms (Koster, 2004; Arampatzis et al., 2000). So, it becomes necessary to develop a conflation mechanism able to project all the semantically equivalent forms of a phrase into the same index term (Kelledy and Smeaton, 1997; Galvez et al., 2006). In this work, we have opted for a mid-level representation, half-way between an easy-to-compute plain representation, and a more complete, but also more complex, syntactic representation such as trees (Smeaton et al., 1995; Sheridan and Smeaton, 1992; Vilares et al., 2001b, 2004b) or graphs (Montes-y-Gómez et al., 2000).

Our approach relies on the extraction of the dependencies established between the different words contained in a sentence. When limiting the number of terms involved in a dependency to two words, we will obtain *syntactic dependency pairs.*[2] In particular, we have considered the following syntactic dependencies (Carrol et al., 1998):

- Noun-Adjective, relating the head of a noun phrase with its modifying adjective.
- Noun-Adjective prepositional phrase,[3] relating the head of a noun phrase with the head of the modifying prepositional phrase.
- Subject-Verb, relating the subject head with the main verb of the clause.
- Verb-Object/Adjunct, relating the main verb of the clause with the head of its object or adjunct.

Such dependency pairs will be used as complex index terms, called *head-modifier* pairs (Fagan, 1987; Strzalkowski and Perez-Carballo, 1994; Koster, 2004; Arampatzis et al., 2000). It has to be noted that while the head-modifier relation may suggest semantic dependence, what we obtain here is strictly syntactic, even though the semantic relation is our real target (Mittendorfer and Winiwarter, 2002; Perez-Carballo and Strzalkowski, 2000; Arampatzis et al., 2000; Smeaton et al., 1995; Sheridan and Smeaton, 1992).

Once the pairs have been identified and extracted, their component terms are themselves conflated, firstly by means of lemmatization, and then by morphological families (see Section 1). We are thus eliminating both the inflectional changes associated with syntactic and morpho-syntactic variants and the derivational transformations of morpho-syntactic variants, which involve both syntactic variation and derivational transformations (Jacquemin and Tzoukermann, 1999; Jacquemin, 2001). Some related approaches can be found in the literature. In Koster (2004) such components are lemmatized in order to eliminate inflection. In Strzalkowski and Perez-Carballo (1994), Mitra et al. (1997) and Kelledy and Smeaton (1997), stemming techniques are used instead, also covering derivative phenomena. Finally, in Arampatzis et al. (2000) verb nominalization and noun verbalization are proposed for this task.

In order to extract the dependencies, we must first analyze the syntactic structure of documents and queries. Full parsing (Alonso et al., 1999; Sikkel, 1997) is non-viable because of its high computational cost, which makes its application on a large scale impractical. Moreover, the lack of robustness of such approaches, which seek to obtain a complete parse of the whole sentence, greatly reduces their coverage (Arampatzis et al., 2000). This situation is even more problematic in the case of Spanish, due to the lack of freely available resources such as grammars, treebanks, etc. In this context, seeking to

[2] For example, the phrase *"a big fierce dog"* contains two dependency pairs: between the noun *"dog"* and the modifying adjective *"big"*, and between the noun *"dog"* and the modifying adjective *"fierce"*.

[3] By *adjective prepositional phrase* we mean those prepositional phrases acting as adjectives that modify a noun. For example: *"The girl* ***with blue eyes****"*.

obtain a compromise between the quality of the syntactic information to be extracted and the ease of its extraction, the employment of *shallow parsing* techniques (Abney, 1997) enables us both to reduce computational complexity and increase robustness. Shallow parsing has shown itself to be useful in several NLP application fields, particularly in Information Extraction (Aone et al., 1998; Grishman, 1995; Hobbs et al., 1997). Nevertheless, its application in IR has not yet been studied in depth, previous studies having mainly focused on languages other than Spanish and having often been limited to the obtaining of simple noun phrases (Kraaij and Pohlmann, 1998; Mitra et al., 1997; Hull et al., 1997).

There are also some approaches based on the use of existing terminological databases to extract a lexicalized grammar. For instance, the work of Jacquemin (Jacquemin, 2001; Jacquemin and Tzoukermann, 1999) is based on the term lists extracted from thesauri used for manual indexing at INIST/CNRS, a documentation center for scientific and technical information that produces two bibliographical databases, PASCAL and FRANCIS, indexed with a controlled thesaurus. We have not followed this approach due to the fact that only scarce, small and often non-free resources of this kind are available for Spanish (ACRoTermite, 2007; VERBA, 2007; Buyse, 2003; Crespo León et al., 2005; Husson et al., 2000; Reynoso et al., 2000).[4]

3. Shallow Parsing Through Cascades of Transducers: The Cascade Parser

We have developed an advanced, modular, widely applicable and robust parser, named Cascade, based on cascades of finite-state transducers. The theoretical basis for its design comes from Formal Language Theory (Hopcroft and Ullman, 1979), which tells us that, given a context-free grammar and an input string, the syntactic trees of height k generated by a parser can be obtained by means of k layers of finite-state transducers: the first layer obtains the nodes labeled by non-terminals corresponding to left-hand sides of productions that only contain terminals on their right-hand side; the second layer obtains those nodes which only involve terminal symbols and those non-terminal symbols generated on the previous layer; and so on.

3.1. *System architecture*

The shallow parser is based on a five-layer architecture whose input is the output of a tagger-lemmatizer.[5] The rest of the section describes how each layer works. For this purpose, we will use as our notation context-free rules extended with classical regular expression operators. In the same way, uppercase identifiers denote a set of terms, which can either be pre-terminals, namely tags resulting from part-of-speech tagging, or elements of a given grammatical category. When the presence of a concrete lemma is required, this will be indicated by using the `typewriter` font.

[4] Notice that the lack of freely available linguistic resources has been referred to as one of the motives underlying the work reported in this article.

[5] In particular, we have employed `MrTagoo` (Graña et al., 2001, 2002), although any high-performance part-of-speech tagger could be used.

3.1.1. *Layer 0: Preprocessing Extension*

This layer extends the linguistic preprocessing capability of the system, minimizing the noise generated during the subsequent parsing steps. It deals with:

- *Numerals in non-numerical format.*
- *Quantity expressions.* Expressions of the type *algo más de dos millones* (a little more than two million), which denote a number but with a certain vagueness about its concrete value, are identified as *numeral phrases* (*NumP*).
- *Expressions with a verbal function.* Some verbal expressions such as *tener en cuenta* (to take into account), must be considered as a unit —in this case synonym of the verb *considerar* (to consider)— to avoid errors in the upper layers such as identifying *en cuenta* as an object or adjunct of the verb.

3.1.2. *Layer 1: Adverbial Phrases and First Level Verbal Groups*

This layer consists of rules only containing tags and/or lemmas in its right-hand side. To enable the next layers to extract syntactic dependency pairs, we will associate to the non-terminal in the left side of each rule, firstly, the lemma corresponding to the head of the phrase we are recognizing, and secondly, the tag with the appropriate morphosyntactic features. The notation employed for this inheritance mechanism is inspired in the notation employed when specifying the set of restrictions in feature structure-based grammars (Carpenter, 1992).

The first rule we describe here allows us to identify sequences of adverbs (*W*), called *adverbial phrases* (*AdvP*). The last adverb will be considered the phrase head, so its lemma and its tag will be the lemma and tag of the non-terminal *AdvP*:

$$AdvP \quad \rightarrow \quad W^* \;\; W_1 \quad \begin{cases} AdvP.lem \doteq W_1.lem \\ AdvP.tag \doteq W_1.tag \end{cases}$$

The following set of rules allows us to identify *first level verbal groups* (*VG1*) —or non-periphrastic verbal groups— corresponding to passive forms,[6] whether simple tenses, e.g. *soy observado* (I am observed), or compound tenses,[7] e.g. *he sido observado* (I have been observed), active forms being identified in a similar way. The first rule manages compound forms: the tag is taken from the auxiliary verb `haber` (to have), whereas the lemma is taken from the main verb, which must be a participle, the same as the auxiliary verb `ser` (to be). The second rule manages simple forms: the tag is obtained from the form of the auxiliary verb `ser`, whereas the lemma is taken from the main verb, again a participle.

$$VG1 \quad \rightarrow \quad V_1 \;\; V_2 \;\; V_3 \quad \begin{cases} VG1.lem \doteq V_3.lem \\ VG1.tag \doteq V_1.tag \\ VG1.voice \doteq \textsc{pass} \\ V_1.lem \doteq \texttt{haber} \\ V_2.lem \doteq \texttt{ser} \\ V_2.tense \doteq \textsc{part} \\ V_3.tense \doteq \textsc{part} \end{cases} \qquad VG1 \quad \rightarrow \quad V_1 \;\; V_2 \quad \begin{cases} VG1.lem \doteq V_2.lem \\ VG1.tag \doteq V_1.tag \\ VG1.voice \doteq \textsc{pass} \\ V_1.lem \doteq \texttt{ser} \\ V_2.tense \doteq \textsc{part} \end{cases}$$

[6] Constructed with the auxiliary verb `ser` (to be).

[7] Constructed with the auxiliary verb `haber` (to have).

3.1.3. *Layer 2: Adjectival Phrases and Second Level Verbal Groups*

Adjectival phrases (*AdjP*) and second level verbal groups (*VG2*) are processed here. An adjectival phrase (*AdjP*) —e.g. *azul* (blue) or *muy alto* (very tall)— is formed by a head adjective (*A*) sometimes preceded by an adverbial phrase (*AdvP*) modifying it:

$$AdjP \quad \rightarrow \quad AdvP? \;\; A \quad \begin{cases} AdjP.lem \doteq A.lem \\ AdjP.tag \doteq A.tag \end{cases}$$

Second level verbal groups (*VG2*) include periphrastic verbal forms such as *tengo que ir* (I have to go). *Verbal periphrases* are unions of two or more verbal forms working as a unit, giving to the semantics of the main verb attributive shades of meaning, such as obligation, degree of development of the action, etc., which can not be expressed by means of the simple and compound forms of the verb. A periphrasis is generally formed by a conjugated auxiliary verb giving the inflection, a verb in a non-personal form (infinitive, gerund or participle) giving the main meaning, and an optional element (preposition or conjunction) linking both verbs.

Infinitive periphrases are identified using the following rule, which takes into account the possibility of the existence of an enclitic pronoun (previously separated from the verb form by the tagger) when the auxiliary verb is reflexive. The tag is inherited from the auxiliary verb, while the lemma and the voice are inherited from the main verb:

$$VG2 \quad \rightarrow \quad VG1_1 \;\; (\texttt{me}|\texttt{te}|\texttt{se})? \;\; (\texttt{que}|\texttt{de}|\texttt{a})? \;\; VG1_2 \quad \begin{cases} VG2.lem \doteq VG1_2.lem \\ VG2.tag \doteq VG1_1.tag \\ VG2.voice \doteq VG1_2.voice \\ VG1_1.voice \doteq \textsc{act} \\ VG1_2.tense \doteq \textsc{inf} \end{cases}$$

Gerund and participle periphrases are managed in a similar way, whereas first level verbal groups which do not take part in any periphrastic group are promoted to second level verbal groups.

3.1.4. *Layer 3: Noun Phrases*

Noun phrases (*NP*) are processed in this layer.[8] We have taken into account the possibility of their being preceded by a *partitive complement* (*PC*) such as `alguno de` (some of), `ninguno de` (none of), etc.:[9]

$$\begin{array}{ll} PC \rightarrow (& \texttt{algún} \mid \texttt{alguno} \mid \texttt{cualquier} \mid \texttt{cualquiera} \mid \\ & \texttt{ningún} \mid \texttt{ninguno} \mid \texttt{mucho} \mid \texttt{uno} \;\;)_1 \;\; \texttt{de} \end{array} \quad \begin{cases} PC.num \doteq ()_1.num \end{cases}$$

[8] As has been stated at the beginning of this section, this shallow parser is only able to process syntactic structures of limited depth. So, we are restricted to those *simple* noun phrases formed by a noun head and its adjectival modifiers. If we want to extend the processing to more complex noun phrases also including prepositional modifiers, we would have to extend the cascade by at least one more layer, since *simple* prepositional phrases (formed by *simple* noun phrases introduced by a preposition) are processed in layer 4, as shown in Sect. 3.1.5.

[9] *Partitive complements* denote a part of the whole.

Following the head of the noun phrase, there may appear an adjectival post-modifier consisting of two adjectival phrases coordinated by a conjunction (Cc), or consisting of a sequence of one, two or even three adjectival phrases:

$$AdjPostModif \rightarrow (\ AdjP\ \ Cc\ \ AdjP\ \mid\ AdjP\ \mid\ AdjP\ \ AdjP\ \mid\ AdjP\ \ AdjP\ \ AdjP\ )$$

The head of the noun phrase is formed by a common noun (N), an acronym or a proper noun; its tag and lemma will decide the tag and lemma of the whole phrase. In the case of several candidates for head appearing, we will take the last one. Optionally, we may find one or more determiners (D) and an adjectival phrase before the head.[10] The existence of adjectival post-modifiers is also optional, and thus we finally obtain the rule:

$$\begin{array}{ll} NP \rightarrow PC?\ \ D^*\ \ (AdjP \mid Number \mid NumP)? & \\ \qquad (N \mid Acronym \mid Proper)^*\ \ (N \mid Acronym \mid Proper)_1 & \left\{ \begin{array}{l} NP.lem \doteq ()_1.lem \\ NP.tag \doteq ()_1.tag \\ NP.number \doteq PC.number \end{array} \right. \\ \qquad AdjPostModif? & \end{array}$$

3.1.5. *Layer 4: Prepositional Phrases*

The function of the last layer is to identify *prepositional phrases* (PP, $PPof$, $PPby$), i.e. those formed by a noun phrase (NP) preceded by a preposition (P). To make the extraction of dependencies easier, we will distinguish those phrases introduced by the prepositions `de` (of) and `por` (by) from the rest, producing the following rules:[11]

$$PPof \quad \rightarrow \quad P \quad NP \left\{ \begin{array}{l} P.lem \doteq \texttt{de} \\ PP.lem \doteq NP.lem \\ PP.tag \doteq NP.tag \end{array} \right. \qquad PPby \quad \rightarrow \quad P \quad NP \left\{ \begin{array}{l} P.lem \doteq \texttt{por} \\ PP.lem \doteq NP.lem \\ PP.tag \doteq NP.tag \end{array} \right.$$

$$PP \quad \rightarrow \quad P \quad NP \left\{ \begin{array}{l} PP.lem \doteq NP.lem \\ PP.tag \doteq NP.tag \end{array} \right.$$

3.2. *Identification of Syntactic Roles*

The syntactic roles we are trying to identify, and the heuristics used for this purpose, are the following:

- *Subject.* The closest noun phrase (NP) preceding a verbal group ($VG2$) in personal form will be considered the subject of the sentence.
- *Object.* This is the closest NP after an active non-copula $VG2$.
- *Agentive BY-phrase.*[12] This is the closest $PPby$ following a passive non-copula $VG2$.
- *Subject complement.* For a copula verb, we will identify as the subject complement that non-attached $AdjP$ or that head of a $NP/PPof$ closest to the verbal group.

[10] Spanish allows adjectival modifiers to appear either before or after the modified noun, although the latter is the default order.

[11] Most prepositional phrases are ignored during further processing, except those introduced by the prepositions `de` (of) and `por` (by). This way, by managing them separately, processing becomes simpler.

[12] That is, those prepositional phrases that appear in a passive sentence introduced by `por` (in Spanish: `by` in English) and that become the subject when the sentence is changed into the active mood.

- *Adjunct.* Due to the problem of prepositional phrase attachment, we have opted for a strict criterion when searching them in order to minimize the noise introduced by erroneous identifications. We will only consider as an adjunct that prepositional phrase following the verb which is closest to it, and previous to any subject complement, object or adjunct identified beforehand.
- *Adjective prepositional phrase.* Due to the ambiguity in the attachment of prepositional phrases, we will only consider the prepositional *PPof* phrases due to their high reliability. So, when the system finds a *PPof* immediately after a noun or prepositional phrase, it is identified as an adjective prepositional phrase.

3.3. *Extraction of Dependencies*

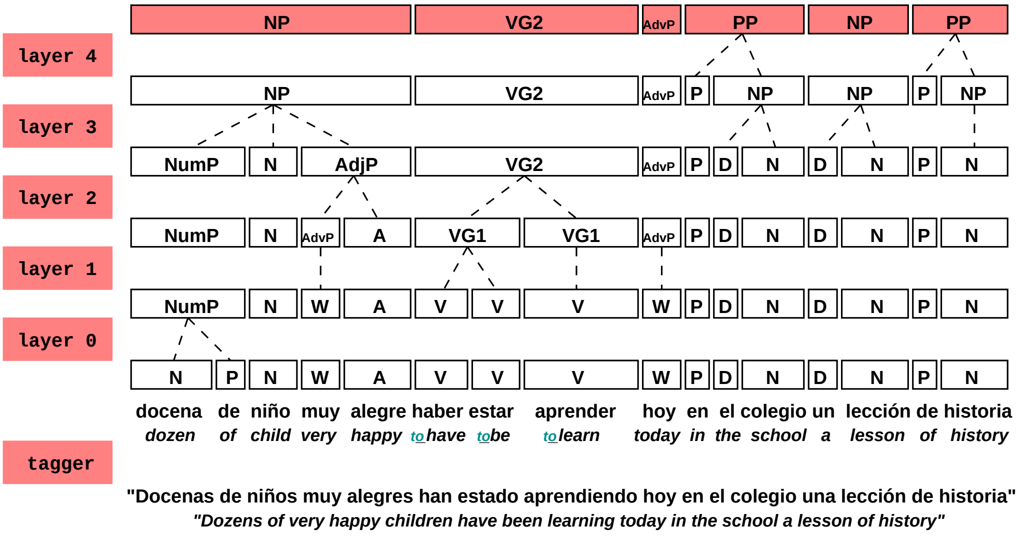
Once we have identified the syntactic roles of the phrases of the sentence, the syntactic dependencies existing between them are extracted in the form of pairs that involve:

- A noun and each of its modifying adjectives. In fact, whereas the rest of dependencies are extracted once the parsing process has finished, dependencies of this kind are extracted during the identification of noun phrases in layer 3. This is because such dependencies are internal to the noun phrase, and therefore, if they are not extracted at that point, the information would be lost once the phrase is reduced to its head.
- A noun and the head of its modifying adjective prepositional phrase.
- The head of the subject and the non-copula verb.
- The head of the subject and the head of the subject complement, since from a semantic point of view copula verbs act as mere links between them.
- An active verb and the head of the object.
- A passive verb and the head of the agentive BY-phrase.
- A non-copula verb and the head of the adjunct.
- The head of the subject and the head of the adjunct, but only when it is a copula sentence, due to the special behavior of copula verbs, as described above.

For each dependency extracted, their components are conflated through both lemmatization and morphological families, as described in Section 2.

3.4. *Implementation of the Shallow Parser*

Each of the rules involved in the different stages of the parsing process has been implemented through a finite-state transducer. Our goal is to obtain, as pairs, a list of the syntactic dependencies of the text. The formation of such pairs only involves the heads of the phrases, so we only need to retain the lemma of the head, together with its corresponding morpho-syntactic features. When extending such an approach to all layers, text will be formed by `lemma tag non-terminal` triplets. One of the main advantages of employing this representation is that it allows us to make references to any of the components (i.e. lemmas, tags and non-terminals) in the right-hand side of the rules. In order to preserve uniformity when initializing the system, we will consider that every grammatical category is a valid non-terminal. This way, at the beginning, the output of the tagger is directly translated into the format required by the parser: `lemma tag non-terminal`. Each time the parser finds a matching of the right-hand side of the rule



layer 4
layer 3
layer 2
layer 1
layer 0
tagger
NP VG2 AdvP PP NP PP
NP VG2 AdvP P NP NP P NP
NumP N AdjP VG2 AdvP P D N D N P N
NumP N AdvP A VG1 VG1 AdvP P D N D N P N
NumP N W A V V V W P D N D N P N
N P N W A V V V W P D N D N P N
docena de niño muy alegre haber estar aprender hoy en el colegio un lección de historia
dozen of child very happy to have to be to learn today in the school a lesson of history
"Docenas de niños muy alegres han estado aprendiendo hoy en el colegio una lección de historia"
"Dozens of very happy children have been learning today in the school a lesson of history"

Figure 1. Overview of the parsing process for the running example.

with the input terms, the rule is reduced, and the matching terms are replaced by the left-hand side of the rule. Usually, the lemma and tag of the phrase are inherited from the lemma and the tag of its head. The output of the parser is a sequence of *lemma* tag non-terminal triplets corresponding to the phrases recognized during the parsing process.

3.5. *A Running Example*

Let us take as example the following sentence: *Docenas de niños muy alegres han estado aprendiendo hoy en el colegio una lección de historia* (Dozens of very happy children have been learning today in the school a lesson of history). The first step consists of tagging and lemmatizing the input sentence —you are reminded that, in this initial stage, the non-terminal is the grammatical category of the term: [13]

[*docena (dozen)* NCFP *N*] [*de (of)* P *P*] [*niño (child)* NCMP *N*] [*muy (very)* WQ *W*]
[*alegre (happy)* AQFP *A*] [*haber (to have)* V3PRI *V*] [*estar (to be)* VPMS *V*]
[*aprender (to learn)* VRG *V*] [*hoy (today)* WI *W*] [*en (in)* P *P*] [*el (the)* DAMS *DA*]
[*colegio (school)* NCMS *N*] [*un (a)* DAFS *DA*] [*lección (lesson)* NCFS *N*] [*de (of)* P *P*]
[*historia (history)* NCFS *N*]

A summary of the parsing process is shown in Fig. 1. At the output of the cascade we obtain the sequence of heads corresponding to the phrases identified during this process:

[*niño (child)* NCMP *NP*] [*aprender (to learn)* V3PRI *VG2*] [*hoy (today)* WI *AdvP*]
[*colegio (school)* NCMS *PP*] [*lección (lesson)* NCFS *N*] [*historia (history)* NCFS *PPof*]

[13] For a better understanding of the example, we will complete this notation by adding the translation of the lemma and by separating the terms by means of square brackets.

The dependencies contained in the parsed text are then extracted in order to be conflated into complex index terms. Firstly, according to the criteria established in Section 3.2, the syntactic roles of the phrases are identified. In the case of our example, an active subject (*SUBJact*), an active non-copula verbal group (*Vact*), its adjunct (*Adjunct*), its object (*OBJ*), and the adjective prepositional phrase of the latter (*APP*) are identified:

[*niño (child)*	NCMP	*NP*]	–	⟨ *SUBJact* ⟩
[*aprender (to learn)*	V3PRI	*VG2*]	–	⟨ *Vact* ⟩
[*hoy (today)*	WI	*AdvP*]	–	⟨ ⟩
[*colegio (school)*	NCMS	*PP*]	–	⟨ *Adjunct* ⟩
[*lección (lesson)*	NCFS	*NP*]	–	⟨ *OBJ* ⟩
[*historia (history)*	NCFS	*PPof*]	–	⟨ *APP* ⟩

Next, the dependencies indicated in Section 3.3 are extracted, obtaining as output:

ADJ	(*niño*	NCMP,	*alegre*	AQFP)
APP	(*lección*	NCFS,	*historia*	NCFS)
SUBJact	(*aprender*	V3PRI,	*niño*	NCMP)
OBJ	(*aprender*	V3PRI,	*lección*	NCFS)
Adjunct	(*aprender*	V3PRI,	*colegio*	NCMS)

The tests performed with this new parser have shown a reliable behavior when identifying the existence of syntactic dependencies between two words. Nevertheless, such dependencies are not always correctly classified. This is the case, for example of subjects coming after verbs, which are identified as objects or adjuncts by the heuristics.[14] The existence of a dependency has however been correctly identified. Errors of this kind when identifying the syntactic role of a phrase are not a problem in our IR task, since such information is dismissed when conflating and adding the dependency to the index. The main point here is to detect the dependency correctly, a task that the parser performs reliably.

4. Indexing

Complex terms have to be considered as a complement of simple terms, since the exclusive use of complex terms as index terms, as in the case of the exclusive use of simple terms, enables only a partial and insufficient view of the semantics of the text to be captured (Mitra et al., 1997). Furthermore, when only complex terms are used, system recall is clearly reduced because of the high degree of sparseness of their term space. This is because the number of dependency pairs existing in a collection is much higher than the number of words it contains, since given a set of words, the number of phrases that can be built with them is much higher than the number of words forming such a set. As an example, we show in Table 1 the distribution of both word lemmas (*lem*) and dependency pairs (*qdp*) in the Spanish CLEF 2003 corpus (CLEF, 2007),

[14] In Spanish, the order of sentence and phrase constituents is less strict than in English. Constructions not following the default *subject-verb-object* word order are usual in common language.

Table 1
Distribution, by document frequency (*df*), of the lemmas and complex terms extracted from the test corpus CLEF 2003 using lemmatization (*lem*), shallow parsing (*qdp*), and shallow parsing restricted to noun phrases (*qnp*)

	lem		***qdp***		***qnp***	
df	***#lemmas***	***%lemmas***	***#pairs***	***%pairs***	***#pairs***	***%pairs***
[1..1]	216 403	51.76%	2 227 774	56.91%	1 171 381	57.04%
[2..2]	56 524	13.52%	591 121	15.10%	311 022	15.14%
[3..4]	42 330	10.12%	425 179	10.86%	221 998	10.81%
[5..8]	30 345	7.26%	280 921	7.18%	145 041	7.06%
[9..16]	21 026	5.03%	173 262	4.43%	88 544	4.31%
[17..32]	14 652	3.50%	103 181	2.64%	53 491	2.60%
[33..∞)	36 805	8.81%	112 995	2.89%	62 253	3.03%
Total	418 085	100.00%	3 914 433	100.00%	2 053 730	100.00%
***df*>4**	102 828	24.59%	670 359	17.13%	349 329	17.01%

one of the test collections we have used. So, the probability of occurrence of the same phrase in two different documents is much lower than that of the words it contains and consequently the probability of a matching during the retrieval process is also much lower. This is why complex terms should be used in combination with simple terms (Narita and Ogawa, 2000; Mitra et al., 1997; Hull et al., 1997; Strzalkowski and Perez-Carballo, 1994; Buckley et al., 1993; Sheridan and Smeaton, 1992; Buckley et al., 1993; Fagan, 1987). In this work, dependency pairs are used as index terms together with the content word lemmas of the text. However, the combined use of simple and complex terms creates several problems:

(i) The assumption of term independence is violated since words forming a pair also occur in the documents from which the dependency has been extracted (Narita and Ogawa, 2000).

(ii) There is an over-balance of the weight of complex terms, which occur much less frequently than simple terms because of their high degree of sparseness and, therefore, their assigned weight is much higher (Smeaton et al., 1995; Strzalkowski and Perez-Carballo, 1994).

This situation introduces an element of instability into the system, because when undesired matchings of complex terms with non-relevant documents occur, their relevances increase excessively. At the same time, and also due to the same reason, when correct matchings between complex terms and relevant documents occur, we obtain a clear improvement of the results with respect to the employment of simple terms only. It can be argued that according to this we would expect similar results to those obtained only with simple terms. Nevertheless, complex term matchings are much less frequent than those for simple terms because of their high sparseness. Therefore, fortuitous matchings of complex terms are much more harmful than those for simple terms, whose effect tends to be weakened by the rest of the matchings. Thus, we can state that the noise intro-

Table 2
Evaluation corpora: composition of the document subcollections (showing their size, the number of documents they contain, and the average length of those documents in number of words)

subcollection	*size (MB)*	*#docs.*	*avg. length (words)*
EFE 1994	509	215 738	317.64
EFE 1995	577	238 307	325.33
EFE 1994+1995	1086	454 045	321.67

duced by erroneous complex term matchings is amplified. We therefore need to solve this over-balance of complex terms in order to minimize the negative effect of undesired matchings.

The solution to both problems consists in decreasing the extra initial relevance assigned to complex terms by introducing a balance factor between the weights of simple and complex terms (Narita and Ogawa, 2000; Hull et al., 1997).

5. Evaluation

In order to evaluate our approach, it has been integrated into the well-known vector-based engine SMART (Buckley, 1985), using an `atn-ntc` weighting scheme (Savoy, 2003). Since our aim is to investigate whether syntactic processing can be used to improve the performance of classic IR systems, we have chosen as working environment a classic configuration which can be considered, to a certain extent, standard.

The system has been tested using the CLEF 2001–2003 Spanish monolingual test collection, [15] which has become the standard evaluation corpus for Spanish IR tasks. An IR test collection is composed of three parts (Baeza-Yates and Ribeiro-Neto, 1999): the documents, the example information requests —called *topics* in the literature—, and a list of relevant documents for each of these topics.

In this case, the document collection is formed by news reports, formatted in SGML, provided by the Spanish news agency EFE. The initial collection, used in CLEF 2001–2002, was formed by news reports from the year 1994 (subcollection EFE 1994) and was enlarged in CLEF 2003 with news reports from the year 1995 (EFE 1995). The composition of these subcollections is shown in Table 2. [16]

With respect to the topic set, 50 topics were used in CLEF 2001 (numbered 41 to 90), a further 50 in CLEF 2002 (91 to 140) and 60 in CLEF 2003 (141 to 200). The topics are formed by three fields, as shown in Fig. 2: a brief *title* statement, a one-sentence *description*, and a more complex *narrative* specifying the relevance assessment criteria. Nevertheless, only title and description fields have been used here for generating the test queries. The resulting ”short” queries are, to a certain extent, an acceptable approximation to those queries used in commercial engines, and have used for this purpose

[15] CLEF: Cross-Language Evaluation Forum, an organization for the promotion of research in cross-language IR that organizes an annual conference, for which the corpus was created and from which it takes its name (CLEF, 2007).

[16] It should be noted that news reports are often written with no care, contain errors and misspellings (Figuerola et al., 2001), and therefore have a negative impact in NLP approaches.

```
<top>
<num> C071 </num>
<ES-title> Verduras, frutas y cáncer </ES-title>
<ES-desc> Encontrar documentos que relacionen la ingestión de verduras y fruta con el
cáncer. </ES-desc>
<ES-narr> Son relevantes aquellos documentos que informen de los efectos positivos o
negativos de ingerir fruta y verduras sobre el cáncer. </ES-narr>
</top>


<top>
<num> C071 </num>
<EN-title> Vegetables, Fruit and Cancer </EN-title>
<EN-desc> Find documents that relate the eating of vegetables and fruit to cancer.
</EN-desc>
<EN-narr> Documents reporting either positive or negative effects of eating fruit and
vegetables on cancer are relevant. </EN-narr>
</top>
```

Figure 2. Sample topic: topic number 71 in Spanish (top) and English (bottom)

Table 3
Evaluation corpora: final composition of the corpora used (showing the document collections used, the number of topics —notice that only *title* and *description* fields have been used—, and the average length of the resulting topics (*title+description*) in number of words)

corpus	***collection***	***#topics***	***avg. length (words)***
CLEF 2001-02·A	EFE 1994	46	19.28
CLEF 2001-02·B	EFE 1994	45	20.24
CLEF 2003	EFE 1994+1995	47	21.31

by the research community repeatedly (Hull, 1996). Moreover, experiments using these fields are those required by the CLEF organization for the official workshop ranking.

On the other hand, following Hull et al. (1997), those topics with less than five relevant documents have been removed. This is because when the number of relevant documents is too small, any minor variation in the ranking of just one or two documents can involve noticeable changes in the results for that query, thus distorting the global results.

Finally, in order to maximize the homogeneity of the evaluation corpus, queries from CLEF 2001 and 2002 were combined.[17] Three corpora, whose statistics are shown in Table 3, resulted from this process:

CLEF 2001-02·A: a corpus for training and parameter estimation formed by the CLEF 2001–2002 document collection and the odd-numbered topics from these editions.

CLEF 2001-02·B: a corpus for evaluation employing the same news collection as the previous one, but using the even topics.

CLEF 2003: a corpus for evaluation composed of the full set of CLEF 2003 documents and topics.

[17] It should be taken into account that CLEF 2001 was the first year in which Spanish was used. The team responsible for Spanish therefore had no previous experience.

We will now show the results obtained when using syntactic dependencies as complex index terms for complementing simple terms. Firstly, Section 5.1 discusses the results of the tests performed using the syntactic information extracted from queries. Next, Section 5.2 discusses the results obtained when the syntactic information extracted from documents was used.[18]

5.1. *Results Using the Syntactic Information Extracted from Queries*

5.1.1. *Results for All Dependencies*

In this first set of experiments we combine lemmatized simple terms and complex terms obtained from queries. The terms extracted from documents, both simple and complex, are combined and added to the index as described in Section 4. Query processing is similar.

The weight of simple terms is multiplied by a balance factor ω in order to reduce the relative contribution of complex terms in a ratio $1/\omega$. This factor ω has been calculated using the training corpus CLEF 2001-02·A. The following values —and their corresponding ratios— were considered:

$$\omega \in \{1\ \mathit{(1)},\ 2\ \mathit{(0.500)},\ 3\ \mathit{(0.333)},\ 4\ \mathit{(0.250)},\ 5\ \mathit{(0.200)},\ 8\ \mathit{(0.125)},\ 10\ \mathit{(0.100)},\\ 12\ \mathit{(0.083)},\ 14\ \mathit{(0.071)},\ 16\ \mathit{(0.062)},\ 18\ \mathit{(0.055)},\ 20\ \mathit{(0.050)}\}$$

After this tuning process, a factor ω=8 (*ratio* $1/\omega$=*0.125*) was chosen.

Table 4 shows the results obtained applying the balance factor when using the pairs extracted with CASCADE (*qdp*), compared with respect to our baseline, the stemming of simple terms (*stm*),[19] those results for which we have obtained positive improvement appearing in boldface. Column $\underset{stm}{\%\Delta}$ shows the percentage of improvement attained. In order to highlight the improvement strictly due to the use of dependencies, column $\underset{lem}{\%\Delta}$ shows the degree of improvement of *qdp* with respect to lemmatization only.[20]

Each row of the results table contains one of the parameters employed to measure performance (Baeza-Yates and Ribeiro-Neto, 1999): number of queries submitted, number of documents retrieved, number of relevant documents expected, number of relevant documents retrieved, average precision (non-interpolated) for all relevant documents (averaged over queries), R-precision, precision at 11 standard recall levels, and precision at N documents retrieved.

[18] Note that these results must be considered as *non-official* from a strict CLEF point of view, since they have not been evaluated by the CLEF organization.

[19] We have used the Snowball Spanish stemmer (`http://snowball.tartarus.org`), based on Porter's algorithm (Porter, 1980) and one of the most popular stemmers in the IR research community. The stopword list used was that provided by the University of Neuchatel (`http://www.unine.ch/info/clef/`), also commonly used in research. Following the work of Mittendorfer and Winiwarter (2002, 2001), a second list of so-named *meta-stopwords* has been also used for queries. Such stopwords correspond to metalevel content, i.e. those expressions corresponding to query formulation without giving any useful information for the search. This is the case, for example, of the phrase *"encuentre aquellos documentos que describan ... "* (find those documents describing ...).

[20] In the case of lemmatization, index terms are formed by those lemmas of the content words of the text (nouns, adjectives and verbs), the grammatical categories which concentrate the semantics of a text. The corresponding stopword lists consist of the lemmas of the content words of the original lists.

Table 4
Results obtained through stemming (*stm*), baseline, and syntactic dependency pairs obtained from the query with the shallow parser Cascade (*qdp*)

corpus	***CLEF 2001-02·A***				***CLEF 2001-02·B***				***CLEF 2003***			
approach	***stm***	***qdp***	***%*** Δ_{stm}	***%*** Δ_{lem}	***stm***	***qdp***	***%*** Δ_{stm}	***%*** Δ_{lem}	***stm***	***qdp***	***%*** Δ_{stm}	***%*** Δ_{lem}
#queries	46	46	–	–	45	45	–	–	47	47	–	–
#docs. retr.	46k	46k	–	–	45k	45k	–	–	47k	47k	–	–
#rel. exp.	3007	3007	–	–	2513	2513	–	–	2335	2335	–	–
#rel. retr.	2719	**2728**	**0.33**	**1.04**	2345	**2371**	**1.11**	**0.34**	2137	**2169**	**1.50**	**0.46**
Non-int. pr.	.4720	**.4965**	**5.19**	**2.82**	.4674	**.4810**	**2.91**	**2.82**	.4304	**.4377**	**1.70**	**1.23**
R-pr.	.4599	**.4895**	**6.44**	**0.97**	.4584	**.4672**	**1.92**	**2.70**	.4479	**.4490**	**0.25**	**2.93**
Pr. at 0%	.8443	.8406	-0.44	**1.36**	.8645	.8137	-5.88	-1.18	.7881	**.8409**	**6.70**	**3.51**
Pr. at 10%	.7361	**.7640**	**3.79**	**2.37**	.6937	**.7128**	**2.75**	**4.13**	.7002	.6993	-0.13	-0.91
Pr. at 20%	.6377	**.6925**	**8.59**	**2.27**	.6280	**.6608**	**5.22**	**5.11**	.6099	**.6364**	**4.34**	**4.02**
Pr. at 30%	.5769	**.6338**	**9.86**	**3.26**	.5723	**.6137**	**7.23**	**5.28**	.5610	.5566	-0.78	**1.14**
Pr. at 40%	.5351	**.5622**	**5.06**	**2.18**	.5394	**.5575**	**3.36**	**0.36**	.4932	**.4992**	**1.22**	-1.32
Pr. at 50%	.4812	**.5076**	**5.49**	**2.94**	.4979	**.5232**	**5.08**	**1.87**	.4464	**.4625**	**3.61**	-0.69
Pr. at 60%	.4489	**.4648**	**3.54**	**3.38**	.4356	**.4552**	**4.50**	**3.15**	.3961	**.4103**	**3.58**	**2.14**
Pr. at 70%	.3776	**.3909**	**3.52**	**1.45**	.4078	.4002	-1.86	**1.50**	.3378	**.3385**	**0.21**	-1.08
Pr. at 80%	.3246	**.3364**	**3.64**	**2.65**	.3126	**.3280**	**4.93**	**5.06**	.2652	**.2758**	**4.00**	**2.15**
Pr. at 90%	.2410	**.2483**	**3.03**	**5.39**	.2362	.2352	-0.42	**1.42**	.1909	**.1913**	**0.21**	**5.05**
Pr. at 100%	.1169	**.1197**	**2.40**	0.00	.1158	**.1242**	**7.25**	**2.73**	.1008	.0992	-1.59	**6.44**
Pr. at 5	.6391	**.6913**	**8.17**	**4.60**	.6089	**.6178**	**1.46**	**3.73**	.5745	.5745	0.00	-2.87
Pr. at 10	.5935	**.6500**	**9.52**	**3.45**	.5400	**.5822**	**7.81**	**2.74**	.5426	.5128	-5.49	-0.41
Pr. at 15	.5551	**.6029**	**8.61**	**1.70**	.5081	**.5333**	**4.96**	**3.15**	.4908	.4794	-2.32	**1.18**
Pr. at 20	.5174	**.5620**	**8.62**	**3.20**	.4878	**.5078**	**4.10**	**4.10**	.4468	.4436	-0.72	-0.47
Pr. at 30	.4710	**.5036**	**6.92**	**2.19**	.4422	**.4526**	**2.35**	**1.66**	.3986	**.4007**	**0.53**	**0.17**
Pr. at 100	.3157	**.3348**	**6.05**	**1.45**	.2922	**.3024**	**3.49**	**1.04**	.2477	**.2506**	**1.17**	-0.28
Pr. at 200	.2186	**.2263**	**3.52**	**1.30**	.1979	**.2004**	**1.26**	**0.35**	.1611	**.1633**	**1.37**	**0.99**
Pr. at 500	.1097	**.1103**	**0.55**	**1.19**	.0980	**.0992**	**1.22**	**0.92**	.0813	**.0833**	**2.46**	**0.73**
Pr. at 1000	.0591	**.0593**	**0.34**	**1.02**	.0521	**.0527**	**1.15**	**0.38**	.0455	**.0461**	**1.32**	**0.44**

Results are certainly positive since there is clear improvement at all levels, although the behavior of corpus CLEF 2003 is more irregular. Our results are also qualitatively better than those of Mittendorfer and Winiwarter (2002) for English or those of Kraaij and Pohlmann (1998) for Dutch, since there is not only an improvement in precision at the first documents retrieved —as in their case—, but also a general improvement with regard to global precisions.

When taking stemming (*stm*) as the baseline, the Wilcoxon test (α=0.05) shows a significant improvement in corpus CLEF 2001-02·A when comparing non-interpolated precisions, while no significant improvement is shown for the other two corpora. In the case of comparing precision at the top 10 documents, significant improvement is found in the two CLEF 2001-02 corpora. On the other hand, when taking lemmatization (*lem*) as the baseline, the Wilcoxon test obtains the same results, except in the case of corpus CLEF 2001-02·B, where pairs (*qdp*) perform significantly better regarding non-interpolated precision, but not in the case of precision at the top 10 documents.[21]

We also show in column *qdp* of Table 1 the distribution, by document frequency (*df*), of the complex terms extracted from the CLEF 2003 corpus using Cascade. These statistics include the number and percentage of pairs in each frequency rank, together with the total number of pairs contained in the index —those pairs occurring in more than four documents (*df*>4). Figures show that only a small proportion of the pairs is added to the index (17%) because of the high sparseness of the term space. Moreover, the number of unique pairs generated by Cascade is 43% lower than that of the number of pairs generated by other approaches (Vilares et al., 2002), this having a major impact on the size of the index created. Such a difference is due to the inherent conservatism of Cascade when dealing with noun-adjective prepositional phrase dependencies, allowing the attachment of a prepositional phrase with the noun on its left only for *de* (of) prepositions, since it is unable to accurately disambiguate the dependency structure in the case of other prepositions —the classical prepositional phrase attachment problem. Other parsers are more permissive, allowing the attachment of any prepositional phrase to the noun on its left, thus boosting the number of dependencies generated, even when a high percentage of them are incorrect dependencies that introduce noise into the system.

5.1.2. *Results for Noun Phrase Dependencies*

In order to facilitate the comparison of our work with those classical approaches based on the use of noun phrases as index terms (Hearst et al., 1996), a new set of experiments was performed, this time restricted to those dependencies corresponding to noun phrases: those between a noun head and its modifying adjectives, and those between a noun head and its modifying adjective prepositional phrases.

Although the Wilcoxon test showed no significant difference with when all dependencies are used, except for precision at the top 10 documents for CLEF 2003, the results obtained through this approach[22] showed a clear downward trend with regard to the previous one due to the loss of the information contained in dependencies involving verb phrases. The document frequency distribution of the terms generated is shown in column *qnp* of Table 1. As can be seen, the distribution remains almost the same as with the complete set of dependencies —column *qdp*—, but the number of unique terms in the index is 48% lower, clearly reducing the size of the index. It should be noted that although in this way the costs associated with storage and management of the index can be reduced, term generation costs remain the same, since noun and prepositional phrases are identified in layers 3 and 4, respectively, while verb groups, which are necessary for the rest of the dependencies, are processed in the previous layers.

[21] It can be seen that, in general, our approach shows a less improvement when compared to lemmatization than when compared to stemming. This is because, as explained in Section 1, lemmatization performs better than stemming in the case of Spanish (Vilares et al., 2004a).

[22] With a balance factor re-tuned at ω=5 (*ratio* $1/\omega$=*0.200*) after eliminating dependencies involving verb phrases.

5.2. *Results Using the Syntactic Information Extracted from Documents*

Previous experiments showed that the use of syntactic information enables us to obtain a consistent improvement with regard to the use of simple terms only. Nevertheless, we were of the opinion that these results could be improved by applying a more sophisticated mechanism for extracting syntactic dependencies. To this end, we shifted our focus from the syntactic information of queries to the syntactic information of documents, employing a blind feedback-based approach in which the indexing process remains the same, but the querying process is performed in three stages:

(i) The lemmatized query is submitted to the system.

(ii) The n top documents retrieved by this initial query are used to select the most informative dependencies in order to expand the lemmatized query, but with no re-weighting. These dependencies are selected automatically from the t' best terms (both lemmas and dependencies) of the top n'_1 documents using Rocchio's approach to feedback (Rocchio, 1971).

(iii) The expanded query is then submitted to the system in order to obtain the final set of documents retrieved.

The work of Buckley et al. (1993) also refers to the expansion of the queries with the X best phrases of the top documents retrieved (apart from the Y best simple terms); nevertheless, there are clear differences with our approach: the former takes the X best phrases, whereas our approach takes the best dependencies among the X best terms; the former does not use linguistic phrases, but rather pairs of adjacent non-stopwords; Buckley et al.'s approach is applied on *routing* tasks, whereas our proposal is applied on *ad-hoc* retrieval tasks; the relevance criteria they are using for feedback has been set manually instead of applying blind feedback, as in our case, where we are just assuming that the top n'_1 documents retrieved are relevant, and therefore their selected documents are much more reliable than ours when making feedback; and finally, although both approaches use SMART, the weighting scheme they employed was much simpler and had a worse baseline performance, resulting in a greater margin for improvement.

5.2.1. *Results for All Dependencies*

Before evaluation, we first needed to tune the parameters of the model, not only the weight balance factor ω, as usual, but also the expansion parameters: the number of terms t' to extract and the number of top documents n'_1 to assume as relevant. This tuning process was performed with the training corpus CLEF 2001-02·A for the following ranges of parameters n'_1 and t':

$$n'_1 \in \{5, 10, 15, 20\} \qquad\qquad t' \in \{5, 10, 15, 20, 30, 40, 50\}$$

where the resulting parameters were:

$$n'_1{=}10,\ t'{=}50,\ \omega{=}3\ \textit{(ratio } 1/\omega\textit{=0.333)}$$

The results obtained with this new approach employing syntactic information from documents (ddp) are shown in Table 5. As can be seen, there is a generalized improvement, not only with regard to stemming (stm) —see column $\underset{stm}{\%\Delta}$—, but also with regard to query dependency pairs (qdp) —see column $\underset{qdp}{\%\Delta}$—, suggesting that syntactic information from documents is more useful than that from queries when it comes to increasing

Table 5
Results obtained through stemming (*stm*), baseline, and syntactic dependency pairs obtained from the documents (*ddp*)

corpus	***CLEF 2001-02·A***				***CLEF 2001-02·B***				***CLEF 2003***			
approach	***stm***	***ddp***	$\%\Delta_{stm}$	$\%\Delta_{qdp}$	***stm***	***ddp***	$\%\Delta_{stm}$	$\%\Delta_{qdp}$	***stm***	***ddp***	$\%\Delta_{stm}$	$\%\Delta_{qdp}$
#queries	46	46	–	–	45	45	–	–	47	47	–	–
#docs. retr.	46k	46k	–	–	45k	45k	–	–	47k	47k	–	–
#rel. exp.	3007	3007	–	–	2513	2513	–	–	2335	2335	–	–
#rel. retr.	2719	**2758**	**1.43**	**1.10**	2345	**2394**	**2.09**	**0.97**	2137	**2246**	**5.10**	**3.55**
Non-int. pr.	.4720	**.5286**	**11.99**	**6.47**	.4674	**.4990**	**6.76**	**3.74**	.4304	**.4690**	**8.97**	**7.15**
R-pr.	.4599	**.5119**	**11.31**	**4.58**	.4584	**.4765**	**3.95**	**1.99**	.4479	**.4582**	**2.30**	**2.05**
Pr. at 0%	.8443	.8389	-0.64	-0.20	.8645	.8033	-7.08	-1.28	.7881	**.7964**	**1.05**	-5.29
Pr. at 10%	.7361	**.7794**	**5.88**	**2.02**	.6937	**.6990**	**0.76**	-1.94	.7002	**.7300**	**4.26**	**4.39**
Pr. at 20%	.6377	**.7244**	**13.60**	**4.61**	.6280	**.6681**	**6.39**	**1.10**	.6099	**.6564**	**7.62**	**3.14**
Pr. at 30%	.5769	**.6497**	**12.62**	**2.51**	.5723	**.6224**	**8.75**	**1.42**	.5610	**.5818**	**3.71**	**4.53**
Pr. at 40%	.5351	**.5926**	**10.75**	**5.41**	.5394	**.5860**	**8.64**	**5.11**	.4932	**.5499**	**11.50**	**10.16**
Pr. at 50%	.4812	**.5534**	**15.00**	**9.02**	.4979	**.5535**	**11.17**	**5.79**	.4464	**.5046**	**13.04**	**9.10**
Pr. at 60%	.4489	**.5005**	**11.49**	**7.68**	.4356	**.4779**	**9.71**	**4.99**	.3961	**.4439**	**12.07**	**8.19**
Pr. at 70%	.3776	**.4343**	**15.02**	**11.10**	.4078	**.4371**	**7.18**	**9.22**	.3378	**.3959**	**17.20**	**16.96**
Pr. at 80%	.3246	**.3678**	**13.31**	**9.33**	.3126	**.3406**	**8.96**	**3.84**	.2652	**.3166**	**19.38**	**14.79**
Pr. at 90%	.2410	**.2775**	**15.15**	**11.76**	.2362	**.2581**	**9.27**	**9.74**	.1909	**.2303**	**20.64**	**20.39**
Pr. at 100%	.1169	**.1522**	**30.20**	**27.15**	.1158	**.1434**	**23.83**	**15.46**	.1008	**.1100**	**9.13**	**10.89**
Pr. at 5	.6391	**.7000**	**9.53**	**1.26**	.6089	**.6533**	**7.29**	**5.75**	.5745	**.6128**	**6.67**	**6.67**
Pr. at 10	.5935	**.6717**	**13.18**	**3.34**	.5400	**.5800**	**7.41**	-0.38	.5426	**.5745**	**5.88**	**12.03**
Pr. at 15	.5551	**.6203**	**11.75**	**2.89**	.5081	**.5363**	**5.55**	**0.56**	.4908	**.5390**	**9.82**	**12.43**
Pr. at 20	.5174	**.5935**	**14.71**	**5.60**	.4878	**.5056**	**3.65**	-0.43	.4468	**.4957**	**10.94**	**11.74**
Pr. at 30	.4710	**.5348**	**13.55**	**6.20**	.4422	**.4607**	**4.18**	**1.79**	.3986	**.4468**	**12.09**	**11.50**
Pr. at 100	.3157	**.3474**	**10.04**	**3.76**	.2922	**.3078**	**5.34**	**1.79**	.2477	**.2719**	**9.77**	**8.50**
Pr. at 200	.2186	**.2336**	**6.86**	**3.23**	.1979	**.2049**	**3.54**	**2.25**	.1611	**.1739**	**7.95**	**6.49**
Pr. at 500	.1097	**.1117**	**1.82**	**1.27**	.0980	**.1002**	**2.24**	**1.01**	.0813	**.0875**	**7.63**	**5.04**
Pr. at 1000	.0591	**.0600**	**1.52**	**1.18**	.0521	**.0532**	**2.11**	**0.95**	.0455	**.0478**	**5.05**	**3.69**

the precision of the documents retrieved.

When using stemming (*stm*) as the baseline, the Wilcoxon test shows significant improvement for corpora CLEF 2001-02·A and 2003 regarding non-interpolated precision, and also for precision at the top 10 documents in the case of corpus CLEF 2001-02·A. When query dependency pairs (*qdp*) are used as the baseline, the same results are obtained for non-interpolated precision, but this time only significant improvement in precision at the top 10 documents is found for corpus CLEF 2003.

5.2.2. *Results for Noun Phrase Dependencies*

The positive results obtained with the new approach led us to investigate the possibility of the existence of some kind of relation between the index terms introduced by each

Table 6
Number of complex index terms extracted from queries (*qdp*) and documents (*ddp*)

corpus	$qdp\backslash ddp$	$ddp\backslash qdp$	$ddp{\cap}qdp$
CLEF 2001-02·A	106	267	38 (26.38%)
CLEF 2001-02·B	121	291	42 (25.76%)
CLEF 2001-03	126	286	48 (27.58%)

Table 7
Percentage distribution, according to the type of dependency associated, of the complex index terms obtained from queries (*qdp*) and those in common with those obtained from documents (*ddp*∩*qdp*)

corpus	***CLEF 2001-02·A***		***CLEF 2001-02·B***		***CLEF 2003***	
dependency	*qdp*	*ddp*∩*qdp*	*qdp*	*ddp*∩*qdp*	*qdp*	*ddp*∩*qdp*
noun–adjective	37.50%	57.89%	34.35%	50.00%	27.01%	31.25%
noun–adj. prep. phrase	45.83%	34.21%	44.78%	45.23%	47.70%	64.58%
subject–active verb	1.38%	0.00%	4.90%	2.38%	6.89%	2.08%
verb–object	6.25%	2.63%	9.20%	7.14%	7.47%	0.00%
subject–passive verb	0.69%	0.00%	0.00%	0.00%	0.00%	0.00%
passive verb–agent. BY-phrase	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
verb–adjunct	9.02%	5.26%	7.36%	2.38%	10.91%	4.16%
subject–subject compl.	0.00%	0.00%	1.22%	0.00%	0.57%	0.00%

approach. Table 6 shows the number of complex terms dismissed when using the syntactic information from documents ($qdp\backslash ddp$), the number of new terms introduced by documents ($ddp\backslash qdp$), and the number of terms common to both approaches ($ddp{\cap}qdp$), together with the percentage of the original query terms this represents. It can be seen that the proportion of common pairs remains nearly constant, which may suggest the existence of some kind of underlying relation according to which only certain kinds of pairs are useful from the point of view of automatic extraction.

Thus, we analyzed the distribution of the different types of syntactic dependencies from which each complex index term has been obtained. The results obtained are shown in Table 7: column *qdp* shows the number of pairs corresponding to each dependency type which were extracted from queries in our initial approach, whereas column *ddp*∩*qdp* shows the number of pairs common to those extracted from the documents.[23] Our aim was to discover whether there was any bias or preference and, as can be seen, dependencies corresponding to noun phrases —*noun–adjective* and *noun–adjective prepositional phrase*— seem to be preferred.

[23] It should be taken into account that the same pair can be simultaneously generated from different types of dependencies, being counted for each of them.

Since some kind of preference would appear to exist in the case of noun phrase dependencies, we decided to study the behavior of the system when using only complex terms obtained from noun dependencies, in a similar way as in the case of the pairs obtained from queries.

The results obtained for this approach,[24] whilst still positive since they continued to outperform stemming, were not as good as those obtained when the whole set of dependencies was used (*ddp*). Moreover, the difference in the figures obtained was greater than that obtained when using the syntactic information from queries. On the other hand, no significant difference between either approach was found when comparing both non-interpolated precision and precision at top 10 documents.

Khan and Khor (2004) also used relevance feedback for selecting relevant noun phrases. However, our approach is more complete since we deal with all kinds of dependency, and also because our evaluation is performed on a large set of standard queries instead of on an ad-hoc set, as in their case.

6. Conclusions and Future Work

Throughout this article we have described the application of phrase-level analysis techniques in order to obtain, on the one hand, more precise and descriptive index terms, and on the other, to manage syntactic variation. For this purpose we have tested an approach based on the use of syntactic dependencies as complex index terms for complementing simple index terms.

We should point out that although some related works have been done for other languages, with English to the fore, Romance languages, and Spanish in particular, have stayed in the background. Furthermore, our proposal integrates numerous features which serve to differentiate it from previous works.

This is the case, for example, of the introduction of mechanisms based on morphological families for the management of derivational morphology, which enables the extension of the processing of strict syntactic variation to morpho-syntactic variation. Moreover, the use of different sources of syntactic information, namely queries and documents, has been also studied, the latter proving more effective, as has the restriction of the dependencies employed to those obtained from noun phrases in order to reduce costs.

In order to minimize the computational cost of the system and at the same time to increase its robustness, syntactic dependencies are obtained through shallow parsing by means of Cascade, a parser based on a cascade of finite-state transducers for emulating full parsing. Since it widely integrates finite-state technology, its computational complexity is linear with respect to the length of the texts to be processed, thus allowing the rapid processing of large collections and their application in practical environments.

Furthermore, we have had to face one of the main problems in non-English Natural Language Processing research, the lack of freely available linguistic resources. The solution for minimizing this problem consisted in restricting the complexity of the solutions proposed by focusing on the employment of lexical information, which is easier to obtain.

[24] With n'_1=10, t'=40 and ω=3 *(ratio* 1/ω*=0.333).*

With regard to our future work, the results presented in this work represent a starting point for the further development of NLP-based approaches to Romance language IR in general, and Spanish IR in particular. However, we still need to maintain our efforts in order to reduce the gap with IR in English as far as possible.

The work presented here opens the door to the possibility of using *selection restrictions* (Gamallo et al., 2005, 2001) for improving the syntactic disambiguation capability of the system, particularly with respect to the prepositional phrase attachment problem, the goal being to increase system recall without damaging precision.

This work also provides a solid basis for the application of mechanisms for dealing with semantic variation. Current approaches for managing semantic variation mainly work at word level, using WordNet (Miller et al., 1990) as their knowledge source (Arampatzis et al., 2000). Such techniques are very sensitive to word-sense ambiguity, requiring the employment of high-performance word-sense disambiguation techniques (Stevenson, 2003; Stokoe et al., 2003). Even so, real improvements are usually obtained only for incomplete and relatively short queries (Voorhees, 1994). One possibility consists in working with phrase-level semantic variation (Jacquemin, 1999), which reduces the problem of word-sense ambiguity because of the existence of a context in the complex term itself. Moreover, it would be possible to employ a fuzzy notion of synonymy which measures the degree of synonymy between two terms (Sobrino et al., 2006). The existence of such a measure allows the establishment of thresholds for query expansion and term weighting during the matching of synonyms, in this way opening up new possibilities for study.

Finally, we are currently studying the possibility of adapting the *local context analysis* approach proposed by Xu and Croft (1996). This technique proved to be less sensitive to the noise introduced by non-relevant documents than blind relevance feedback, which could be useful during the querying process when using the syntactic information extracted from documents.

Acknowledgments

This research has been partially funded by *Ministerio de Educación y Ciencia* and *FEDER* (TIN2004-07246-C03 and HUM2007-66607-C04) and *Xunta de Galicia* (PGIDIT07SIN005206PR, PGIDIT05PXIC-10501PN, PGIDIT05PXIC30501PN and the *Rede Galega de Procesamento da Linguaxe e Recuperación de Información*).

Biographies of the Authors

Jesús Vilares graduated in Computer Science Engineering from Univ. of A Coruña (Spain) in 2000. After a short period as a lecturer in the Univ. of Vigo (Spain), in 2005 he obtained a PhD. in Computer Science from the Univ. of A Coruña. He is currently an assistant professor at this university. His research work focuses on Natural Language Processing, Information Retrieval and Extraction and Question Answering.

Miguel A. Alonso obtained his Bachelor's degree and PhD. in Computer Science in 1993 and 2000, respectively, from Univ. of A Coruña (Spain), where he is currently an associate professor. His research work focuses on the application of Natural Language Processing techniques to Information Retrieval, Information Extraction, and the parsing of Tree Adjoining Grammars and related frameworks.

Manuel Vilares has an MSc. in Applied Mathematics from the Univ. of Santiago de Compostela (Spain, 1987), a MSc. in Software Engineering from CERICS (France, 1988), and a PhD. in Computer Science from Univ. of Nice–Sophia-Antipolis (France, 1992). He initially worked at the INRIA institute (France) and later in Spain (1992), where he became full professor in Computer Science at Univ. of Vigo (2002). His current research work focuses on Natural Language Processing, Logic Programming, Programming Language Design and Information Extraction.

References

ACRoTermite – Terminology of telecommunications database. International Telecommunication Union. http://www.itu.int/terminology/index.html (visited on August 2007).

CLEF. http://www.clef-campaign.org (visited on August 2007).

VERBA – Polytechnic and Plurilingual Terminological Database. European Language Resources Association (ELRA). http://www.elra.info/ (visited on August 2007).

Abney, S., 1997. Partial parsing via finite-state cascades. Natural Language Engineering 2 (4), 337–344.

Alonso, M., Cabrero, D., de la Clergerie, E., Vilares, M., 1999. Tabular algorithms for TAG parsing. In: Proc. of the 9th conference of the European chapter of the ACL (EACL'99). pp. 150–157.

Aone, C., Halverson, L., Hampton, T., Ramos-Santacruz, M., 1998. SRA: Description of the IE^2 system used for MUC-7. In: Proc. of the 7th message understanding conference (MUC-7).

Arampatzis, A., van der Weide, T. P., van Bommel, P., Koster, C., 2000. Linguistically-motivated information retrieval. In: Encyclopedia of Library and Information Science. Vol. 69. Marcel Dekker, Inc, pp. 201–222.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison Wesley and ACM Press.

Buckley, C., 1985. Implementation of the SMART information retrieval system. Tech. rep., Department of Computer Science, Cornell University, source code available at ftp://ftp.cs.cornell.edu/pub/smart (visited on August 2007).

Buckley, C., Allan, J., Salton, G., 1993. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In: Harman, D. K. (Ed.), Proc. of the 2nd text retrieval conference (TREC-2). pp. 45–56.

Buyse, K., 2003. Generating corpora and lexicons for language specific purposes. Experiences from the ElektraVoc-II project. In: Proc. of the 36th International Meeting of the Societas Linguistica Europaea.

Carpenter, B., 1992. The logic of typed feature structures. No. 32 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Carrol, J., Briscoe, T., Sanfilippo, A., 1998. Parser evaluation: A survey and a new proposal. In: Proc. of the 1st international conference on language resources and evaluation (LREC 1998). pp. 447–454.

Crespo León, F., Gutiérrez Díez, F., Rodríguez Ferri, F., León Vizcaíno, L., Cuello Gijón, F., Gimeno, E. J., Zepeda Sein, C., Sánchez Vizcaíno Rodríguez, J. M., Cerón Madrigal, J. J., Cantos Gómez, P., Schudel, A., 2005. The translation into Spanish of the OIE Manual of diagnostic tests and vaccines for terrestrial animals (mammals, birds and bees): problems, solutions and conclusions. Revue Scientifique et technique (International Office of Epizootics) 24 (3), 1095–1104.

Dillon, M., Gray, A., 1983. FASIT: A fully automatic syntactically based indexing system. Journal of the American Society for Information Science 34 (2), 99–108.

Fagan, J. L., 1987. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods (PhD. thesis). Tech. Rep. TR87-868, Cornell University, USA.

Figuerola, C. G., Gómez, R., Zazo Rodríguez, A. F., Alonso Berrocal, J. L., 2001. Stemming in Spanish: A first approach to its impact on information retrieval. In: Peters, C. (Ed.), 2001. Results of the CLEF 2001 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2001 Workshop. pp 197–202.

Galvez, C., de Moya-Anegón, F., Solana, V. H., 2006. Term conflation methods in information retrieval: Non-linguistic and linguistic approaches. Journal of Documentation 61 (4), 520–547.

Gamallo, P., Agustini, A., Lopes, G. P., 2001. Selection restrictions acquisition from corpora. In: The 10th Portuguese Conference on Artificial Intelligence (EPIA'01). Lecture Notes in Artificial Intelligence. Springer-Verlag, pp. 30–43.

Gamallo, P., Agustini, A., Lopes, G. P., 2005. Clustering syntactic positions with similar semantic requirements. Journal of Computational Linguistics. 31 (1), 107–146.

Graña, J., Alonso, M. A., Vilares, M., 2002. A common solution for tokenization and part-of-speech tagging: One-pass Viterbi algorithm vs. iterative approaches. In: Sojka, P., Kopeček, I., Pala, K. (Eds.), Text, Speech and Dialogue. Vol. 2448 of Lecture Notes in Computer Science. Springer-Verlag, pp. 3–10.

Graña, J., Barcala, F. M., Vilares, J., 2002. Formal methods of tokenization for part-of-speech tagging. In: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing. Vol. 2276 of Lecture Notes in Computer Science. Springer-Verlag, pp. 240–249.

Graña, J., Chappelier, J.-C., Vilares, M., 2001. Integrating external dictionaries into stochastic part-of-speech taggers. In: Angelova, G., Bontcheva, K., Mitkov, R., Nocolov, N., Nikolov, N. (Eds.), Proc. of the EuroConference Recent Advances in Natural Language Processing (RANLP 2001). pp. 122–128.

Grishman, R., 1995. The NYU system for MUC-6 or where's the syntax? In: Proc. of the 6th message understanding conference (MUC-6). Morgan Kaufmann Publishers, pp. 167–176.

Hearst, M., Pedersen, J., Pirolli, P., Schutze, H., Grefenstette, G., Hull, D., 1996. Xerox site report: Four TREC-4 tracks. In: Harman (Ed.), Proc. of the 4th text retrieval conference (TREC-4). pp. 97–119.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., Tyson, M., 1997. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In: Roche, E., Schabes, Y. (Eds.), Finite-State Language Processing. MIT Press, pp. 383–406.

Hopcroft, J. E., Ullman, J. D., 1979. Introduction to automata theory, languages and computation. Series in Computer Science. Addison-Wesley.

Hull, D. A., 1996. Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science 47 (1), 70–84.

Hull, D. A., Grefenstette, G., Schulze, B. M., Gaussier, E., Schütze, H., Pedersen, J. O., 1997. Xerox TREC-5 site report: Routing, filtering, NLP, and Spanish tracks. In: Voorhees, E. M., Harman, D. K. (Eds.), Proc. of the 5th text retrieval conference (TREC-5). pp. 167–180.

Husson, J. L., Viscogliosi, N., Romary, L., Descotte, S., Campenhoudt, M. V., 2000. DHYDRO: a generic environment developed to edit and access multilingual terminological data on the Internet. In: Proc. of the Second Conference on Maritime Terminology. pp. 47–61.

Jacquemin, C., 1999. Syntagmatic and paradigmatic representations of term variation. In: Proc. of the 37th annual meeting of the ACL (ACL'99). pp. 341–348.

Jacquemin, C., 2001. Spotting and discovering terms through natural language processing. The MIT Press.

Jacquemin, C., Tzoukermann, E., 1999. NLP for term variant extraction: Synergy between morphology, lexicon and syntax. In: Strzalkowski (1999), pp. 25–74.

Jurafsky, D., Martin, J. H., 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall.

Kelledy, F., Smeaton, A. F., 1997. Automatic phrase recognition and extraction from text. In: Proceedings of 19th Annual BCS-IRSG Colloquium on IR. Workshops in Computing. BCS.

Khan, M. S., Khor, S., 2004. Enhanced web document retrieval using automatic query expansion. Journal of the American Society for Information Science and Technology 55 (1), 29–40.

Koster, C. H. A., 2004. Head/modifier frames for information retrieval. In: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing. Vol. 2945 of Lecture Notes in Computer Science. Springer-Verlag, pp. 420–432.

Kowalski, G., 1997. Information retrieval systems: Theory and implementation. The Kluwer International Series on Information Retrieval. Kluwer Academic Publishers.

Kraaij, W., Pohlmann, R., 1998. Comparing the effect of syntactic vs. statistical phrase indexing strategies for Dutch. In: Nicolaou, C., Stephanidis, C. (Eds.), Research and Advanced Technology for Digital Libraries. Vol. 1513 of Lecture Notes in Computer Science. Springer-Verlag, pp. 605–614.

Manning, C. D., Schütze, H., 1999. Foundations of statistical natural language processing. The MIT Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J., 1990. Introduction to WordNet: An on-line lexical database. International Journal of Lexicography 3 (4), 235–244.

Mitra, M., Buckley, C., Singhal, A., Cardie, C., 1997. An analysis of statistical and syntactic phrases. In: Proc. of the 5th international conference "Recherche d'information assistee par ordinateur" (RIAO-97). pp. 200–214.

Mittendorfer, M., Winiwarter, W., 2001. A simple way of improving traditional IR methods by structuring queries. In: Proc. of the 2001 IEEE international workshop on natural language processing and knowledge engineering (NLPKE 2001).

Mittendorfer, M., Winiwarter, W., 2002. Exploiting syntactic analysis of queries for information retrieval. Data & Knowledge Engineering 42 (3), 315–325.

Montes-y-Gómez, M., López-López, A., Gelbukh, A., 2000. Information retrieval with conceptual graph matching. In: Ibrahim, M., Küng, J., Revell, N. (Eds.), Database and Expert Systems Applications. Vol. 1873 of Lecture Notes in Computer Science. Springer-Verlag, pp. 312–321.

Narita, M., Ogawa, Y., 2000. The use of phrases from query texts in information retrieval. In: Proc. of the 23rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'00). pp. 318–320.

Perez-Carballo, J., Strzalkowski, T., 2000. Natural language information retrieval: Progress report. Information Processing and Management 36 (1), 155–178.

Porter, M. F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Reynoso, G. A., March, A. D., Berra, C. M., Strobietto, R. P., Barani, M., Iubatti, M., Chiaradio, M. P., Serebrisky, D., Kahn, A., Vaccarezza, O. A., Leguiza, J. L., Ceitlin, M., Luna, D. R., Bernaldo de Quiros, F. G., Otegui, M. I., Puga, M. C., Vallejos, M., 2000. Development of the Spanish version of the Systematized Nomenclature of Medicine: methodology and main issues. In: Proc. of the 2000 American Medical Informatics Association Symposium (AMIA). pp. 694–698.

Rocchio, J., 1971. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Ch. Relevance feedback in information retrieval, pp. 313–323.

Savoy, J., 2003. Report on CLEF 2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval. In: Peters, C., Borri, F. (Eds.). Results of the CLEF 2003 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2003 Workshop. pp. 179–188.

Sheridan, P., Smeaton, A. F., 1992. The application of morpho-syntactic language processing to effective phrase matching. Information Processing and Management 28 (3), 349–369.

Sikkel, K., 1997. Parsing schemata: A framework for specification and analysis of parsing algorithms. Texts in Theoretical Computer Science — An EATCS Series. Springer-Verlag.

Smeaton, A. F., O'Donnell, R., Kelledy, F., 1995. Indexing structures derived from syntax in TREC-3: System description. In: NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC 3). pp. 55–63.

Sobrino, A., Fernández-Lanza, S., Graña, J., 2006. Access to a large dictionary of Spanish synonyms: A tool for fuzzy information retrieval. In: Herrera-Viedma, E., Pasi, G., Crestani, F. (Eds.), Soft computing in web information retrieval: Models and applications. Vol. 197 of Studies in Fuzziness and Soft Computing. Springer-Verlag, pp. 299–316.

Stevenson, M., 2003. Word sense disambiguation: The case for combinations of knowledge sources. Studies in Computational Linguistics. CSLI.

Stokoe, C., Oakes, M. P., Tait, J., 2003. Word sense disambiguation in information retrieval revisited. In: Proc. of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'03). pp. 159–166.

Strzalkowski, T. (Ed.), 1999. Natural Language Information Retrieval. Vol. 7 of Text, Speech and Language Technology. Kluwer Academic Publishers.

Strzalkowski, T., Perez-Carballo, J., 1994. Recent developments in natural language text retrieval. In: Harman, D. K. (Ed.), Proc. of the 2nd text retrieval conference (TREC-2). pp. 123–136.

Tzoukermann, E., Klavans, J., Jacquemin, C., 1997. Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsing. In: Proc. of the 20th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'97). pp. 148–155.

Vilares, J., Alonso, M. A., Vilares, M., 2004a. Morphological and syntactic processing for text retrieval. In: Galindo, F., Takizawa, M., Traunmüller, R. (Eds.), Database and Expert Systems Applications. Vol. 3180 of Lecture Notes in Computer Science. Springer-Verlag pp. 371–380.

Vilares, J., Cabrero, D., Alonso, M. A., 2001a. Applying productive derivational morphology to term indexing of Spanish texts. In: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing. Vol. 2004 of Lecture Notes in Computer Science. Springer-Verlag, pp. 336–348.

Vilares, M., Ribadas, F. J., Graña, J., 2001b. Approximately common patterns in shared-forests. In: Paques, H., Liu, L., Grossman, D. (Eds.), Proc. of the 2001 ACM CIKM – 10th international conference on information and knowledge management. pp. 73–80.

Vilares, M., Ribadas, F. J., Vilares, J., 2004b. Phrase similarity through the edit distance. In: Galindo, F., Takizawa, M., Traunmüller, R. (Eds.), Database and Expert Systems Applications. Vol. 3180 of Lecture Notes in Computer Science. Springer-Verlag, pp. 306–317.

Vilares, J., Barcala, F. M., Alonso, M. A., 2002. Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish. In: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing. Vol. 2276 of Lecture Notes in Computer Science. Springer-Verlag, pp. 381–390.

Voorhees, E. M., Jul. 1994. Query expansion using lexical-semantic relations. In: Proc. of the 17th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'94). pp. 61–69.

Xu, J., Croft, W. B., 1996. Query expansion using local and global document analysis. In: Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96). pp. 4–11.