ABSTRACT OF PhD THESIS

Automatic Semantic Role Labeling using Selectional Preferences with Very Large Corpora

Determinación Automática de Roles Semánticos usando Preferencias de Selección sobre Corpus muy Grandes

Graduated: Hiram Calvo
Center for Research in Computing (CIC)
National Polytechnic Institute (IPN)
Mexico City, Mexico, 07738
hcalvo@cic.ipn.mx
hiramcalvo@gmail.com
Graduated on June 19th, 2006

Advisor: Dr. Alexander Gelbukh
Computing Research Center (CIC)
National Polytechnic Institute (IPN)
Mexico City, Mexico, 07738
www.gelbukh.com

Abstract
We present a method for recognizing semantic roles for Spanish sentences. This method is based on dependency parsing using heuristic rules to infer dependency relationships between words, and word co-occurrence statistics (learnt in an unsupervised manner) to resolve ambiguities such as prepositional phrase attachment. If a complete parse cannot be produced, a partial structure is built with some (if not all) dependency relations identified. Evaluation shows that in spite of its simplicity, the parser's accuracy is superior to the available existing parsers for Spanish. Though certain grammar rules, as well as the lexical resources used, are specific for Spanish, the suggested approach is language-independent. A particularly interesting ambiguity which we have decided to analyze deeper, is the Prepositional Phrase Attachment Disambiguation.
The system uses an ordered set of simple heuristic rules for determining iteratively the relationships between words to which a governor has not been yet assigned. For resolving certain cases of ambiguity we use co-occurrence statistics of words collected previously in an unsupervised manner, whether it be from big corpora, or from the Web (through a search engine such as Google). Collecting these statistics is done by using Selectional Preferences.
In order to evaluate our system, we developed a Method for Converting a Gold Standard from a constituent format to a dependency format. Additionally, each one of the modules of the system (Selectional Preferences Acquisition and Prepositional Phrase Attachment Disambiguation), is evaluated in a separate and independent way to verify that they work properly. Finally we present some Applications of our system: Word Sense Disambiguation and Linguistic Steganography.
Keywords: dependency parsing, pp attachment disambiguation, constituent to dependency conversion, heuristic rules, hybrid parser, selectional preferences.

Resumen
Se presenta un método para reconocer los roles semánticos de las oraciones en español, es decir, identificar el papel que tiene cada uno de los elementos de la oración. Este método se basa en análisis de dependencias usando reglas heurísticas para inferir relaciones de dependencia entre palabras, así como estadísticas de co-ocurrencia (aprendidas de manera no supervisada) para resolver ambigüedades como la adjunción de sintagma preposicional. Si no se puede producir un análisis completo, se construye una estructura parcial con algunas (si no todas) relaciones de dependencia identificadas. La evaluación muestra que a pesar de su simplicidad, la precisión del analizador es superior a aquella de los analizadores existentes actuales para el español. A pesar de que ciertas reglas gramaticales y los recursos léxicos usados son específicos para el español, el enfoque sugerido es independiente del lenguaje. Una ambigüedad interesante que hemos decidido analizar a mayor profundidad, es la desambiguación de sintagma preposicional.

Computación y Sistemas Vol. 12 No. 1, 2008, pp 128-150
ISSN 1405-5546

El sistema usa un conjunto ordenado de reglas heurísticas simples para determinar iterativamente las relaciones entre palabras para las cuales no se les ha asignado aún un gobernante. Para resolver ciertos casos de ambigüedad usamos estadísticas de co-ocurrencias de palabras. Estas estadísticas han sido obtenidas previamente de una manera no supervisada, ya sea a partir de grandes corpus de texto, o a través de Internet (a través de un motor de búsqueda como Google). El conjunto de estadísticas de co-ocurrencias de uso conforman una base de datos de Preferencias de Selección.

Para evaluar este sistema, desarrollamos un método para convertir un estándar existente, de un formato de constituyentes a un formato de dependencias. Adicionalmente, cada uno de los módulos del sistema (Adquisición de Preferencias de Selección, Desambiguación de Sintagma Preposicional) se evalúa de una forma separada e independiente para verificar su correcto funcionamiento. Finalmente, presentamos algunas aplicaciones de nuestro sistema: Desambiguación de sentidos de palabras y Estaganografía lingüística.

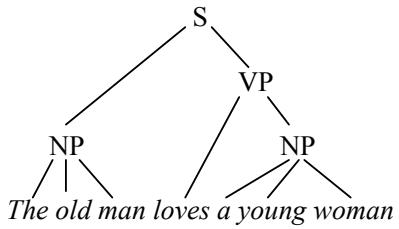
Palabras clave: análisis de dependencias, desambiguación de frase preposicional, conversión de constituyentes a dependencias, reglas heurísticas, analizador sintáctico híbrido, preferencias de selección.

1 Introduction

The two main approaches to syntactic pattern analysis are those oriented to the constituency and dependency structure, respectively. In the constituency approach, the structure of the sentence is described by grouping words together and specifying the type of each group, usually according to its main word [20]:

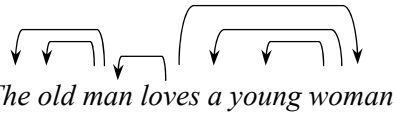
$$[[\text{The old man}]_{NP}\ [\text{loves}\ [\text{a young woman}]_{NP}]_{VP}]_{S}$$

Here NP stands for noun phrase, VP for verb phrase, and S for the whole sentence. Such a tree can also be represented graphically:

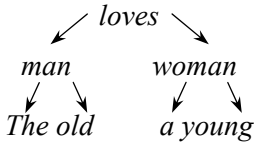


where the nodes stand for text spans (constituents) and arcs for "consists of" relationship.

In dependency approach, words are considered "dependent" from, or modifying, other words [148]. A word modifies another word (governor) in the sentence if it adds details to the latter, while the whole combination inherits the syntactic (and semantic) properties of the governor: *old man* is a kind of *man* (and not a kind of *old*); *man loves woman* is a kind of (situation of) *love* (and not, say, a kind of *woman*). Such dependency is represented by an arrow from the governor to the governed word:



or, in a graphical form:



where the arcs represent the dependency relation between individual words, the words of the lower levels contributing details to those of the upper levels while preserving the syntactic properties of the latter.

In spite of the 40-year discussion in literature, there is no consensus as to which formalism is better. Though combined formalisms such as HPSG [47] have been proposed, they seem to bear the heritage of the advantages as well as disadvantages of both approaches, the latter impeding their wide use in natural language processing practice. Probably the pertinence of each approach depends on a specific task.

We had two-fold motivation for this work. One task we had in mind was the study of lexical compatibility of specific words, and in particular, compilation and use of a dictionary of collocations—stable or frequent word combinations, such as *eat bread* or *deep sleep* as opposed to **eat sleep* and **deep bread* [7]. Such combinations were shown to be useful in tasks ranging from syntactic analysis [58] and machine translation to semantic error correction [8] and steganography [5]. Dependency approach to syntax seems to be much more appropriate for such task.

Our second motivation was the construction of semantic representation of text, even if partial, for a range of applications from information retrieval and text mining [39, 38] to software specifications [24]. All known semantic approaches—such as conceptual graphs [49], Minimal Recursion Semantics [22], or semantic networks [34]—roughly resemble a set of predicates, where individual words represent predicates or their arguments (which in turn can be predicates). The resulting structures are in much closer direct correspondence with the dependency tree than with a constituency tree of the sentence in question, so that dependency syntax seems to be more appropriate for direct translation into semantic structures. Specifically, dependency structure makes it much easier matching—say, in information retrieval—paraphrases of the same meaning (such as active/passive voice transformation) or transforming from one such synonymous structure to another one.

In addition, we found that a dependency parser can be much easier made robust than a constituency parser. The known approaches to dependency parsing cope much easier with both incomplete grammars and ungrammatical sentences than the standard approaches to context-free parsing.

Indeed, a standard context-free parser builds the structure incrementally, so that failure of constructing a constituent implies the impossibility to construct all the further constituents that should have contained this one. What is more, an incorrect decision on an early stage of parsing leads to completely or largely incorrect final result.

In contrast, in dependency parsing the selection of a governor for a given word, or the decision on whether the given two words are connected with a dependency relation, is much more (though not at all completely) decoupled from the corresponding decision on another pair of words. This makes it possible to continue the parsing process even if some of such decisions could not be made successfully. The resulting structure can prove to be incomplete (with some relationships missing) or not completely correct (with some relationships wrongly identified). However, an incorrect decision on a particular pair of words usually does not cause a snowball of cascaded errors at the further steps of parsing.

In this paper we present DILUCT, a simple robust dependency parser for Spanish. Though some specific rules, as well as the lexical resources and the preprocessing tools used, are specific for Spanish, the general framework is language-independent. An online demo and the source code of the system are available online.[1]

The parser uses an ordered set of simple heuristic rules to iteratively determine the dependency relationships between words not yet assigned to a governor. In case of ambiguities of certain types, word co-occurrences statistics gathered in an unsupervised manner from a large corpus or from the Web (through querying a search engine) is used to select the most probable variant. No manually prepared tree-bank is used for training.

We evaluated the parser by counting the number of correctly identified dependency relationships on a relatively small tree-bank. The experiments showed that the accuracy of our system is superior to that of existing Spanish parsers, such as TACAT [18] and Connexor.

2 Related Work

The Dependency approach to syntax was first introduced by Tesnière [53] and further developed by Mel'čuk [34], who extensively used it in his Meaning ⇔ Text Theory [33, 50] in connection to semantic representation as well as with a number of lexical properties of words, including lexical functions [35, 6].

[1] www.likufanele.com/diluct

One of the first serious attempts to construct a dependency parser we are aware about was the syntactic module of the English-Russian machine translation system ETAP [4]. The parsing algorithm consists of two main steps:

1. All individual word pairs with potentially plausible dependency relation are identified.
2. So-called filters remove links incompatible with other identified links.
3. Of the remaining potential links, a subset forming a tree (namely, a projective tree except for certain specific situations) is chosen.

In ETAP, the grammar (a compendium of situations where a dependency relation is potentially plausible) is described in a specially developed specification language describing the patterns to be searched for in the sentence and the actions on constructing the tree that are to be done when such a pattern is found. Both the patters and the actions are expressed in semi-procedural way, using numerous built-in functions (some of which are language-dependent) used by the grammar interpreter. An average pattern–action rule consists of 10–20 lines of tight code. To our knowledge, no statistical information is currently used in the ETAP parser.

Our work is inspired by this approach. However, we made the following main design decisions different from those of ETAP. First, our parser is meant to be much simpler, even if at the cost of inevitable loss of accuracy. Second, we do not rely on complex and detailed lexical recourses. Third, we rely on word co-occurrences statistics, which we believe to compensate for the lack of completeness of the grammar.

Indeed, Yuret [58] has shown that co-occurrence statistics (more precisely, a similar measure that he calls *lexical attraction*) alone can provide enough information for highly accurate dependency parsing, with no hand-made grammar at all. In his algorithm, of all projective trees the one that provides the highest total value of lexical attraction of all connected word pairs is selected. However, his approach relies on huge quantities of training data (though training is unsupervised). In addition, it only can construct projective trees (a tree is called projective if it has no crossing arcs in the graphical representation shown in Section 1).

We believe that a combined approach using both a simple hand-made grammar and word co-occurrence statistics learned in an unsupervised manner from a smaller corpus provides a reasonable compromise between accuracy and practical feasibility.

On the other hand, the mainstream of current research on dependency parsing is oriented to formal grammars [23]. In fact, HPSG [41] was perhaps one of the first successful attempts to—in effect—achieve a dependency structure (necessary for both using lexical information in the parser itself and constructing the semantic representation) by using a combination of constituency and dependency machinery. As we have mentioned, low robustness is a disadvantage of non-heuristically-based approaches.

Of syntactic parsers with realistic coverage available for Spanish we can mention the commercially available XEROX parser[2] and Connexor Machinese Syntax[3] and the freely available parser TACAT.[4] We used the latter two systems to compare their accuracy with that of our system. Only Connexor's system is really dependency-based, relying on the Functional Dependency Grammar formalism [52], the other systems being constituency-based.

3 Algorithm

Following the standard approach, we first pre-process the input text—incl. tokenizing, sentence splitting, tagging, and lemmatizing—and then apply the parsing algorithm

3.1 Preprocessing

Tokenization and sentence splitting: The text is tokenized into words and punctuation marks and split into sentences.

We currently do not distinguish punctuation marks; thus each punctuation mark is substituted with a comma (in the future we will consider different treatment for different punctuation marks).

Two compounds of article and preposition are split: *del* → *de el* 'of the', *al* → *a el* 'to the'.

[2] which used to be on www.xrce.xerox.com/research/mltt/demos/spanish.html

[3] www.connexor.com/demo/syntax.

[4] www.lsi.upc.es/~nlp/freeling/demo.php.

Sentence	Rule
Un(Det) *perro*(N) *grande*(Adj) *ladra* (V)	Det ← N
perro(N) *grande*(Adj) *ladra* (V) ↓ *Un*(Det)	N → Adj
perro(N) *ladra* (V) ↙ ↘ *Un*(Det) *grande*(Adj)	N ← V
ladra (V) ↓ *perro*(N) ↙ ↘ *Un*(Det) *grande*(Adj)	Done

Fig. 1. Applying rules to parse *Un perro grande ladra* 'a big dork barks'

Compound prepositions represented in writing as several words are jointed into one word, for example: *con la intención de* 'in order to', *a lo largo de* 'throughout', etc. Similarly treated are a few adverbial phrases such as *a pesar de* 'in spite of', *de otra manera* 'otherwise', etc., and several pronominal phrases such as *sí mismo* 'itself'. The list of such combination is small (currently including 62 items) and closed.

Tagging: The text is POS-tagged using the TnT tagger [10] trained on the Spanish corpus CLiC-TALP.[5] This tagger has a performance of over 94%.

We also correct some frequent errors of the TnT tagger, for example:

Rule	Example
Det Adj V → Det S V	*el* □*ngles vino* 'the English(man) came'
Det Adj Prep → Det S Prep	*el* □*ngles con* 'the English(man) with'

Lemmatizing: We use a dictionary-based Spanish morphological analyzer [29].[6] In case of ambiguity the variant of the part of speech (POS) reported by the tagger is selected, with the following exceptions:

Tagger predicted	Analyzer found	Example
Adjective	Past participle	*dado* 'given'
Adverb	Present participle	*dando* 'giving'
Noun	Infinitive	*dar* 'to give'

If the analyzer does not give an option in the first column but does give one in the second column, the latter is accepted. If an expected noun, adjective, or participle is not recognized by the analyzer, we try removing a suffix removal, e.g., *flaquito* → *flaco* 'little (and) skinny → skinny.' For this, we try removing a suspected suffix and check whether the word is recognized by the morphological analyzer. Examples of the suffix removal rules are:

Rule	Example
-cita → *-za*	*tacita* → *taza* 'little cup → cup'
-quilla → *-ca*	*chiquilla* → *chica* 'nice girl → girl'

[5] clic.fil.ub.es.

[6] www.Gelbukh.com/agme.

3.2 Rules

Parsing rules are applied to the lemmatized text. Following an approach similar to [4,15], we represent a rule as a sub-graph, e.g., N ← V. Application of a rule consists in the following steps:

1. A substring matching the sequence of the words in the rule is searched for in the sentence.
2. Syntactic relations between the matched words are established according to those specified in the rule.
3. All words that have been assigned a governor by the rule are removed from the sentence in the sense that they do not participate in further comparisons at step 1.

For example, for the sentence *Un perro grande ladra* 'a big dog barks' see Fig. **1**.

As it can be seen from the example, the order of the rule application is important. The rules are ordered; at each iteration of the algorithm, the first applicable rule is applied, and then the algorithm repeats looking for an applicable rule from the first one. The processing stops when no rule can be applied.

Note that one of consequences of such an algorithm is its natural treatment of repeated modifiers. For example, in the phrases *el otro día* 'the other day' or *libro nuevo interesante* 'new interesting book' the two determiners (two adjectives, respectively) will be connected as modifiers to the noun by the same rule Det ← N (N → Adj, respectively) at two successive iterations of the algorithm. We will give additional comments to some rules. Currently our grammar includes the rules sown in Table 1.

Coordinative conjunctions always have been a pain in the neck of dependency formalisms and an argument in favor of constituency approaches. Following the idea of Gladki [30], we represent coordinated words in a constituency-like manner, joining them in a compound quasi-word. In the resulting "tree" we effectively duplicate (or multiply) each arc coming to, or outgoing from, such a special node. For example, a fragment [*John Mary*] ← *speak* (*John and Mary speak*) is interpreted as representing two relationships: *John* ← *speak* and *Mary* ← *speak*; a fragment *merry* ← [*John Mary*] ← *marry* (*Merry John and Mary marry*) yields for dependency pairs: *merry* ← *John* ← *marry* and *merry* ← *Mary* ← *marry*. We should note that currently this machinery is not fully implemented in our system; accordingly, our rules for handling conjunctions are rewriting rules rather than tree construction rules. The first rule forms such a compound quasi-word out of two coordinated nouns if they precede a plural verb. The rule eliminates the conjunction, since in our implementation conjunctions do not participate in the tree structure. Basically what the rule does is to assure that the verb having such a compound subject is plural, i.e., to rule out the interpretation of *John loves Mary and Jack loves Jill* as *John loves* [*Mary and Jack*] *loves Jill*.

3.3 Prepositional Phrase Attachment

This stage is performed after the stage of application of the rules described in the previous section. For any preposition that have not yet been attached to a governor, its compatibility with every noun and every verb in the sentence is evaluated using word co-occurrence statistics (which can be obtained by a simple query to an Internet search engine). The obtained measure is combined with a penalty on the linear distance: the more distant is a potential governor from the preposition in question the less appropriate it is for attachment.

3.4 Heuristics

The heuristics are applied after the stages described in the previous sections. The purpose of the heuristics is to attach the words that were not assigned any governor in the rule application stage.

The system currently uses the following heuristics, which are iteratively applied in this order, in a manner similar to how rules are applied:

1. An unattached *que* 'that, which' is attached to the nearest verb (to the left or to the right of the *que*) that does not have another *que* as its immediate or indirect governor.
2. For an unattached pronoun is attached to the nearest verb that does not have a *que* as its immediate or indirect governor.
3. An unattached N is attached to the most probable verb that does not have a *que* as its immediate or indirect governor. For estimating the probability, an algorithm similar to the one described in the previous section is used. The statistics described in [17] are used.

4. For an unattached verb *v*, the nearest another verb *w* is looked for to the left; if there is no verb to the left, then the nearest one to the right is looked for. If *w* has a *que* as direct or indirect governor, then *v* is attached to this *que*; otherwise it is attached to *w*.
5. An unattached adverb or subordinative conjunction (except for *que*) is attached to the nearest verb (to the left or to the right of the *que*) that does not have another *que* as its immediate or indirect governor.

Table 1. Grammar rules for parsing

Rule	Example
Auxiliary verb system and verb chains	
estar | *andar* ← Ger	*estar comiendo* 'to be eating'
haber | *ser* ← Part	*haber comido* 'to have eaten'
haber ← *estado* ← Ger	*haber estado comiendo* 'have been eating'
ir_{pres} *a* ← Inf	*ir a comer* 'to be going to eat'
ir_{pres} ← Ger ← Inf	*ir queriendo comer* 'keep wanting to eat'
V → *que* → Inf	*tener que comer* 'to have to eat'
V → V	*querer comer* 'to want to eat'
Standard constructions	
Adv ← Adj	*muy alegre* 'very happy'
Det ← N	*un hombre* 'a man'
N → Adj	*hombre alto* 'tall man'
Adj ← N	*gran hombre* 'great man'
V → Adv	*venir tarde* 'come late'
Adv ← V	*perfectamente entender* 'understand perfectly'
Conjunctions (see explanation below)	
N Conj N V(pl) ⇒ [N N] V(pl)	*Juan y María hablan* 'John and Mary speak'
X Conj X ⇒ [X X] (X stands for any)	(*libro*) *nuevo e interesante* 'new and interesting (book)'
Other rules	
N → *que* V	*hombre que habla* 'man that speaks'
que → V	*que habla* 'that speaks'
N X *que* (X stands for any)	*hombre tal que* 'a man such that'; *hombre , que* 'man, which'
Det ← Pron	*otro yo* 'another I'
V → Adj	*sentir triste* 'to feel sad'
N , Adj	*hombre , alto* 'man , tall'
N , N	*hombre , mujer* 'man , woman'
N → Prep → V	*obligación de hablar* 'obligation to speak'
V , V	*comer , dormir* 'eat , sleep'
V Det ← V	*aborrecer el hacer* 'hate doing'

Note that if the sentence contains more than one verb, at the step 4 each verb is attached to some another verb, which can result in a circular dependency. However, this does not harm since such a circular dependency will be broken in the last stage of processing.

3.5 Selection of the Root

The structure constructed at the steps of the algorithm described in the previous sections can be redundant. In particular, it can contain circular dependencies between verbs. The final step of analysis is to select the most appropriate root.

We use the following simple heuristics to select the root. For each node in the obtained digraph, we count the number of other nodes reachable from the given one through a directed path along the arrows. The word that maximizes this number is selected as the root. All its incoming arcs are deleted from the final structure.

4 Evaluation

We present in this section a comparison of our parser against a hand-tagged gold standard. We also compare our parser with two widely known parsers for Spanish. The first one is Connexor Machinese Syntax for Spanish, a dependency parser, and TACAT, a constituency parser.

We have followed the evaluation scheme proposed by [12], which suggests evaluating parsing accuracy based on grammatical relations between lemmatized lexical heads. This scheme is suitable for evaluating dependency parsers and constituency parsers as well, because it considers relations in a tree which are present in both formalisms, for example [Det *car the*] and [DirectObject *drop it*]. For our purposes of evaluation we translate the output of the three parsers and the gold standard into a series of triples including two words and their relationship. Then the triples of the parsers are compared against the triples from the gold standard.

We have chosen the corpus Cast3LB as our gold standard because it is, until now, the only syntactically tagged corpus for Spanish that is widely available. Cast3LB is a corpus consisting of 100,000 words (approx. 3,700 sentences) extracted from two corpora: the CLiCTALP corpus (75,000 words), a balanced corpus containing literary, journalistic, scientific, and other topics; the second corpus was the EFE Spanish news agency (25,000 words) corresponding to year 2000. This corpus was annotated following [21] using the constituency approach, so that we first converted it to a dependency treebank. A rough description of this procedure follows. For details, see [17].

1. Extract patterns from the treebank to form rules. For example, a node called NP with two children, Det and N yields the rule NP → Det N
2. Use heuristics to find the head component of each rule. For example, a noun will always be the head in a rule, except when a verb is present. The head is marked with the @ symbol: NP → Det @N.
3. Use this information to establish the connection between heads of each constituent
4. Extract triples for each dependency relation in the dependency tree-bank.

As an example, consider Table 2. It shows the triples for the sentence taken from Cast3LB. *El más reciente caso de caridad burocratizada es el de los bosnios, niños y adultos.* 'The most recent case of bureaucratized charity is the one about the Bosnian, children and adult.' In some cases the parsers extract additional triples not found in the gold standard.

We extracted 190 random sentences from the 3LB tree-bank and parsed them with Connexor and DILUCT. Precision, recall and F-measure of the different parsers against Cast3LB are as follows.

	Precision	Recall	F-measure
Connexor	0.55	0.38	0.45
DILUCT	0.47	0.55	0.51
TACAT[7]	–	0.30	–

Note that the Connexor parser, though has a rather similar F-measure as our system, is not freely available and of course is not open-source.

[7] Results for TACAT were kindly provided by Jordi Atserias.

5 Application to Prepositional Phrase Attachment Disambiguation

Extracting information automatically from texts for database representation requires previously well-grouped phrases so that entities can be separated adequately. For example in the sentence *See the cat with a telescope*, two different groupings are possible: *See [the cat] [with a telescope]* or *See [the cat with a telescope]*. The first case involves two different entities, while the second case has a single entity. This problem is known in syntactic analysis as prepositional phrase (PP) attachment disambiguation.

Table 2. Triples extracted for the sentence: *El más reciente caso de caridad burocratizada es el de los bosnios, niños y adultos*

Spanish triples	gloss	3LB	Connexor	DILUCT	TACAT
adulto DET el	'the adult'	x			
bosnio DET el	'the bosnian'	x	x	x	
caridad ADJ burocratizado	'bureaucratized charity'	x		x	x
caso ADJ reciente	'recent case'	x		x	x
caso DET el	'the case'	x		x	x
caso PREP de	'case of'	x	x	x	x
de DET el	'of the'	x			x
de SUST adulto	'of adult'	x			
de SUST bosnio	'of bosnian'	x		x	
de SUST caridad	'of charity'	x	x	x	x
de SUST niño	'of children'	x			
niño DET el	'the child'	x			
reciente ADV más	'most recent'	x			x
ser PREP de	'be of'	x		x	x
ser SUST caso	'be case'	x		x	x
recentar SUST caso	'*to recent* case'		x		
caso ADJ más	'case most'			x	
bosnio SUST niño	'bosnian child'			x	
ser SUST adulto	'be adult'			x	
de ,	'of ,'				x
, los	', the'				x
, bosnios	', Bosnian'				x

There are several methods to disambiguate a PP attachment. Earlier methods, e.g. those described in [43, 11], showed that up to 84.5% of accuracy could be achieved using treebank statistics. Kudo and Matsumoto [31] obtained 95.77% accuracy with an algorithm that needed weeks for training, and Lüdtke and Sato [26] achieved 94.9% accuracy requiring only 3 hours for training. These methods require a corpus annotated syntactically with chunk-marks. This kind of corpora is not available for every language, and the cost to build them can be relatively high, considering the number of person-hours that are needed. A method that works with untagged text is presented in [14]. This method has an accuracy of 82.3, it uses the Web as corpus and therefore it can be slow—up to 18 queries are used to resolve a single PP attachment ambiguity, and each preposition + noun pair found in a sentence multiplies this number.

Table 3. Occurrence examples for some verbs in Spanish

Triplet	Literal English translation	Occurrences	% of total verb occurrences
ir a {actividad}	go to {activity}	711	2.41%
ir a {tiempo}	go to {time}	112	0.38%
ir hasta {comida}	go until {food}	1	0.00%
beber {sustancia}	drink {substance}	242	8.12%
beber de {sustancia}	drink of {substance}	106	3.56%
beber con {comida}	drink with {food}	1	0.03%
amar a {agente_causal}	love to {causal_agent}	70	2.77%
amar a {lugar}	love to {place}	12	0.47%
amar a {sustancia}	love to {substance}	2	0.08%

The algorithm presented in [14] is based on the idea that a very big corpus has enough representative terms that allow PP attachment disambiguation. As nowadays it is possible to have locally very big corpora, we ran experiments to explore the possibility of applying such method without the limitation of an Internet connection. We tested with a very big corpus of 161 million words in 61 million sentences. This corpus was obtained on-line from 3 years of publication of 4 newspapers. The results were disappointing—the same algorithm that used the Web as corpus yielding a recall of almost 90% had a recall of only 36% with a precision of almost 67% using the local newspaper corpus.

Table 4. Examples of Semantic Classifications of Nouns

Word	English translation	Classification
rapaz	predatory	activity
rapidez	quickness	activity
rapiña	prey	shape
rancho	ranch	place
raqueta	racket	thing
raquitismo	rickets	activity
rascacielos	skyscraper	activity
rasgo	feature	shape
rastreo	tracking	activity
rastro	track	activity
rata	rat	animal
ratero	robber	causal agent
rato	moment	place
ratón	mouse	animal
raya	boundary	activity
	manta ray	animal
	dash	shape
rayo	ray	activity
raza	race	grouping
razón	reason	attribute
raíz	root	part
reacción	reaction	activity
reactor	reactor	thing
real	real	grouping
realidad	reality	attribute
realismo	realism	shape
realización	realization	activity
realizador	producer	causal agent

Therefore, our hypothesis is that we need to generalize the information contained in the local newspaper corpus to maximize recall and precision. A way for doing this is using selectional preferences: a measure of the probability of a complement to be used for certain verb, based on the semantic classification of the complement. This way, the problem of analyzing *I see the cat with a telescope* can be solved by considering *I see {animal} with {instrument}* instead.

For example, to disambiguate the PP attachment for the *Spanish* sentence *Bebe de la jarra de la cocina* '(he) drinks from the jar of the kitchen' selectional preferences provide information such as *from {place}* is an uncommon complement for the verb *bebe* 'drinks', and thus, the probability of attaching this complement to the verb *bebe*, is low. Therefore, it is attached to the noun *jarra* yielding *Bebe de [la jarra de la cocina]* '(he) drinks [from the jar of the kitchen]'.

Table 3 shows additional occurrence examples for some verbs in Spanish. From this table it can be seen that the verb *ir* 'to go' is mainly used with the complement *a {activity}* 'to {activity}'. Less used combinations have almost zero occurrences, such as *ir hasta {food}* lit. 'go until food'. The verb *amar* 'to love' is often used with the preposition *a* 'to'.

{food}:	breakfast, feast, cereal, beans, milk, etc.
{activity}:	abuse, education, lecture, fishing, hurry, test
{time}:	dawn, history, Thursday, middle age, childhood
{substance}:	alcohol, coal, chocolate, milk, morphine
{name}:	John, Peter, America, China
{causal_agent}:	lawyer, captain, director, intermediary, grandson
{place}:	airport, forest, pit, valley, courtyard, ranch

Fig. 2. Examples of words for categories shown in Table 4

In this section, we propose a method to obtain selectional preferences information such as that shown in Table 3. In Section 5.1, we will discuss briefly related work on selectional preferences. Sections 5.2 to 5.5 explain our method. In Section 5.6, we present an experiment and evaluation of our method applied to PP attachment disambiguation.

5.1 Related work

The terms *selectional constraints* and *selectional preferences* are relatively new, although similar concepts are present in works such as [55] or [25]. One of the earliest works using these terms was [44], where Resnik considered selectional constraints to determine the restrictions that a verb imposes on its object. Selectional constraints have rough values, such as whether an object of certain type can be used with a verb. Selectional preferences are graded and measure, for example, the probability that an object can be used for some verb [45]. Such works use a shallow parsed corpus and a semantic class lexicon to find selectional preferences for word sense disambiguation.

Another work using semantic classes for syntactic disambiguation is [42]. In this work, Prescher *et al.* use an EM-Clustering algorithm to obtain a probabilistic lexicon based in classes. This lexicon is used to disambiguate target words in automatic translation.

A work that particularly uses WordNet classes to resolve PP attachment is [11]. In this work, Brill and Resnik apply the Transformation-Based Error-Driven Learning Model to disambiguate the PP attachment, obtaining an accuracy of 81.8%. This is a supervised algorithm.

5.2 Sources of Noun Semantic Classification

A semantic classification for nouns can be obtained from existing WordNets, using a reduced set of classes corresponding to the unique top-concepts for WordNet nouns described in [36]. These classes are: activity, animal, life_form, phenomenon, thing, causal_agent, place, flora, cognition, process, event, feeling, form, food, state, grouping, substance, attribute, time, part, possession, and motivation. To these unique top-concepts or

beginners, *name* and *quantity* are added. *name* corresponds to capitalized words not found in the semantic dictionary and *quantity* corresponds to numbers.

Since not every word is covered by WordNet and since there is not a WordNet for every language, the semantic classes can be alternatively obtained automatically from Human-Oriented Explanatory Dictionaries. A method for doing this is explained in detail in [13]. Examples of semantic classification of nouns extracted from the human-oriented explanatory dictionary [32] using this method are shown in Table 4 and Figure 2.

5.3 Preparing Sources for Extracting Selectional Preferences

Journals or newspapers are common sources of great amounts of medium to good quality text. However, usually these media exhibit a trend to express several ideas in little space.

This causes sentences to be long and full of subordinate sentences, especially for languages in which an unlimited number of sentences can be nested. Because of this, one of the first problems to be solved is to break a sentence into several sub-sentences. Consider for example the sentence shown in Figure 4—it is a single sentence, extracted from a Spanish newspaper.

We use two kinds of delimiters to separate subordinate sentences: delimiter words and delimiter patterns. Examples of delimiter words are *pues* 'well'*, ya que* 'given that'*, porque* 'because'*, cuando* 'when'*, como* 'as'*, si* 'if'*, □ore so* 'because of that'*, y luego* 'and then'*, con lo cual* 'with which'*, mientras* 'in the meantime'*, con la cual* 'with which' (feminine)*, mientras que* 'while'. Examples of delimiter patterns are shown in Figure 3. These patterns are POS based, so the text was shallow-parsed before applying them.

The sentence in Figure 4 was separated using this simple technique so that each sub-sentence lies in a different row.

5.4 Extracting Selectional Preferences Information

Now that sentences are tagged and separated, our purpose is to find the following syntactic patterns:

1. Verb $_{\text{NEAR}}$ Preposition $_{\text{NEXT\_TO}}$ Noun
2. Verb $_{\text{NEAR}}$ Noun
3. Noun $_{\text{NEAR}}$ Verb
4. Noun $_{\text{NEXT\_TO}}$ Preposition $_{\text{NEXT\_TO}}$ Noun

PREP V ,	CONJ PRON V	CONJ N V
V ADV que	PREP DET que N	PREP DET V
, PRON V	N que V	, N V
V PREP N , N V	, donde	N , que V
V PREP N , N PRON V	N , N	N , CONJ que
V PREP N V	CONJ N N V	N que N PRON V
V de que	CONJ N PRON V	CONJ PRON que V V

Fig. 3. Delimiter patterns V: verb, PREP: preposition, CONJ: conjunction, DET: determiner, N: noun, lowercase are strings of words

Patterns 1 to 3 will be referred henceforth as *verb patterns*. Pattern 4 will be referred as a *noun* or *noun classification pattern.* The $_{\text{NEAR}}$ operator implies that there might be other words in-between. The operator $_{\text{NEXT TO}}$ implies that there are no words in-between. Note that word order is preserved, thus pattern 2 is different of pattern 3. The results of these patterns are stored in a database. For verbs, the lemma is stored. For nouns, its semantic classification, when available through Spanish WordNet, is stored. As a noun may have several semantic classifications, due to, for example, several word senses, a different pattern is stored for each semantic classification. For example, see Table 6. This table shows the information extracted for the sentence of Figure 4.

Once this information is collected, the occurrence of patterns is counted. For example, the last two rows in Table 6, *fin, de, año* and *fin, de, siglo* add 2 of each of the following occurrences: place of cognition, cognition of cognition, event of cognition, time of cognition, place of time, cognition of time, event of time, and time of time. An example of the kind of information that results from this process is shown in Table 3. This information is used then as a measure of the selectional preference that a noun has to a verb or to another noun.

5.5 Experiment and Results

The procedure explained in the previous sections was applied to a corpus of 161 million words comprising more than 3 years of articles from four different Mexican newspapers. It took approximately three days on a Pentium IV PC to obtain 893,278 different selectional preferences for verb patterns (patterns 1 to 3) for 5,387 verb roots, and 55,469 different semantic selectional preferences for noun classification patterns (pattern 4).

In order to evaluate the quality of the selectional preferences obtained, we tested them on the task of PP attachment disambiguation. Consider the first two rows of Table 6, corresponding to the fragment of text *governed by the laws of the market*. This fragment reported two selectional preferences patterns: *govern by {cognition}* and *govern of {activity/thing}*. With the selectional preferences obtained, it is possible to determine automatically the correct PP attachment: values of co-occurrence for *govern of {activity/thing}* and *{cognition} of {activity/thing}* are compared. The highest one sets the attachment.

Formally, to decide if the noun N_2 is attached to its preceding noun N_1 or is attached to the verb V of the local sentence, the values of frequency for these attachments are compared using the following formula [54]:

$$freq\,(X, P, C_2) = \frac{occ\,(X, P, C_2)}{occ\,(X) + occ\,(C_2)},$$

where X can be V, a verb, or C_1, the classification of the first noun N_1. P is a preposition, and C_2 is the classification of the second noun N_2. If $freq(C_1, P, C_2) > freq(V, P, C_2)$, then the attachment is decided to the noun N_1. Otherwise, the attachment is decided to the verb V. The values of $occ(X, P, C_2)$ are the number of occurrences of the corresponding pattern in the corpus. See Table 3 for examples of verb occurrences. Examples of noun classification occurrences taken from the Spanish journal corpus are: *{place} of {cognition}*: 354,213, *{place} with {food}*: 206, *{place} without {flora}*: 21. The values of *occ(X)* are the number of occurrences of the verb or the noun classification in the corpus. For example, for *{place}* the number of occurrences is 2,858,150.

The evaluation was carried on 3 different files of LEXESP corpus [48], containing 10,926 words in 546 sentences. On average, this method achieved a precision of 78.19% and a recall of 76.04%. Details for each file processed are shown in Table 5.

5.6 Evaulation of using PP Attachment Disambiguation using Selectional Preferences

Using selectional preferences for PP attachment disambiguation yielded a precision of 78.19% and a recall of 76.04%. These results are not as good as the ones obtained with other methods, such as an accuracy of 95%. However, this method does not require any costly resource such as an annotated corpus, nor an Internet connection (using the web as corpus); it does not even need the use of a semantic hierarchy (such as WordNet), as the semantic classes are obtained from Human-Oriented Explanatory Dictionaries, as it was discussed in Section 5.2.

Table 5. Results of the PP attachment disambiguation using selectional preferences

file	#sentences	words	average words per sentence	kind of text	precision	recall
n1	252	4,384	17.40	news	80.76%	75.94%
t1	74	1,885	25.47	narrative	73.01%	71.12%
d1	220	4,657	21.17	sports	80.80%	81.08%
total:	546	10,926		**average**:	78.19%	76.04%

Y ahora, cuando
(el mundo) **está** gobernado por (las leyes del mercado), cuando
(lo determinante en la vida) **es**
comprar o
vender, sin
fijarse en <los que
carecen de todo>,
son fácilmente **comprensibles** <las razones de
<la ola de publicidad global que
convenció <a los posibles compradores de servicios y regalos > de que
había (grandes razones) para
celebrar> y
como les **pareciese** poco (el fin de año)
se **lanzaron** a
propagar (el fin del siglo y del milenio)

And now, when
the world **is** governed by market's laws, when
what determines life **is**
to buy or
to sell without
taking into account those that
don't have anything,
easily understandable **are** the reasons for the global publicity wave that

convinced the possible buyers of services and gifts that

there were great reasons to

celebrate, and
as the end of the year **was** not enough for them,
they **launched** themselves
to propagate the end of the century and the millennium

Fig. 4. Example of a very long sentence in a style typically found in journals
( ) surround simple NPs; < > surround NP subordinate clauses, **verbs** are in boldface

We found also that, at least for this task, applying techniques that use the Web as corpus to local corpora reduces the performance of these techniques in more than 50%, even if the local corpora are very big.

In order to improve results for PP attachment disambiguation using selectional preferences, our hypothesis is that instead of using only 25 fixed semantic classes, intermediate classes can be obtained by using a whole hierarchy. In this way, it would be possible to have a flexible particularization for terms commonly used together, i.e. collocations, such as *fin de año* 'end of year', while maintaining the power of generalization. Another point of further developments is to add a WSD module, so that not every semantic classification for a single word is considered, as it was described in Section 5.5.

6 Application to Word Sense Disambiguation

Selectional Preferences are patterns that measure the degree of coupling of an argument (direct object, indirect object and prepositional complements) with a verb. For example, for the verb *to drink*, the direct objects *water*, *juice*, *vodka*, and *milk* are more probable than *bread*, *ideas*, or *grass*.

In order to have a wide coverage of possible complements for a verb, it is necessary to have a very big training corpus, so that every combination of a verb and a complement be found in such a training corpus. However, even for a corpus of hundreds of millions of words, there are word combinations that do not occur in it; sometimes these word combinations are not used very frequently, or sometimes they are used often but they are not seen in certain training corpora.

A solution for this problem is to use word classes. In this case, *water, juice*, *vodka* and *milk* belong to the class of *liquid* and can be associated with the verb *to drink*. However, not all verbs have a single class that is associated with them. For example the verb *to take* can have arguments of many different classes: *take a seat*, *take place*, *take time*, etc. On the other hand, each word can belong to more than one class. This depends not only on the sense of the word, but the main feature that has been taken into account when assigning it to a class. For example, if we consider the color of the objects, *milk* would belong to the class of white objects. If we consider physical properties, it may

belong to the class of fluids or liquids. *Milk* can be *basic_food* too, for example. We can say then that the relevant classification for a word depends both on its use and the classification system being used.

Table 6. Semantic patterns information extracted from Sentence in Figure 4

Words	Literal translation	Pattern
gobernado, *por*, *ley*	governed, by, law	*gobernar*, *por*, cognition
gobernado, *de*, *mercado*	governed, of, market	*gobernar*, *de*, activity thing
es, *en*, *vida*	is, in, life	*ser*, *en*, state life_form causal_agent attribute
convenció, *a*, *comprador*	convinced, to, buyer	*convencer*, *a*, causal_agent
convenció, *de*, *servicio*	convinced, of, service	*convencer*, *de*, activity process possession thing grouping
pareciese, *de*, *año*	may seem, of, year	*parecer*, *de*, cognition time
lanzaron, *de*, *año*	released, of, year	*lanzar*, *de*, cognition time
propagar, *de*, *siglo*	propagate, of, century	*propagar*, *de*, cognition time
propagar, *de*, *milenio*	propagate, of, millennium	*propagar*, *de*, cognition time
ley, *de*, *mercado*	law, of, market	cognition, *de*, activity thing
ola, *de*, *publicidad*	wave, of, publicity	event, *de*, activity cognition
comprador, *de*, *servicio*	buyer, of, service	causal_agent, *de*, activity process possession thing grouping
fin, *de*, *año*	end, of, year	place cognition event time, *de*, cognition time
fin, de, *siglo*	end, of, century	place cognition event time, *de*, cognition time

To find a correlation between the usage of a noun, its sense, and the selectional preferences for the verbs, the following kind of information is needed: (1) Ontological information for a word —a word is not linked to a single class, but a whole hierarchy, and (2) information of the usage of the word in a sentences, given a verb and its specific position in the ontology.

In this application we explore a method to extract selectional preferences that are linked to an ontology. This information is useful to solve several problems following the approach of pattern-based statistical methods combined with knowledge [46].

Table 7 presents an example of the kind of information we obtain with our method. The table shows the values of argument's co-occurrence with the verb for three Spanish verbs using the WordNet hierarchy. These numbers were obtained following the methodology that is described in detail in the Section 6.2. Note that synsets that have

greater chance of being an argument for a verb have a greater value, such as *drink liquid*. In contrast, lower values indicate that a synset is less likely to be an argument for the corresponding verb (v. gr. *Drink reading*, *read food* or *drink surface*). These combinations were found due to mistakes in the training corpus or due to several unrelated senses of a word. For example, *gin* can be also a *trap* that in turn is a *device*. This may lead to **drink device*. When big corpora are used for training, this noise is substantially reduced in contrast with correct patterns, allowing for disambiguation of word senses based on the sentence's main verb.

Table 7. Non-common usages (lower occurrence values) and common usages (higher occurrence values) of word combinations of verb + WordNet synset

verb	synset	Literal English gloss	Weighted occurrences
leer	*fauna*	'read fauna'	0.17
leer	*comida*	'read food'	0.20
leer	*mensaje*	'read message'	27.13
leer	*escrito*	'read writing'	28.03
leer	*objeto_inanimado*	'read inanimate_object'	29.52
leer	*texto*	'read text'	29.75
leer	*artículo*	'read article'	37.20
leer	*libro*	'read book'	41.00
leer	*comunicación*	'read communication'	46.17
leer	*periódico*	'read newspaper'	48.00
leer	*línea*	'read line'	51.50
beber	*superficie*	'drink surface'	0.20
beber	*vertebrado*	'drink vertebrate'	0.20
beber	*lectura*	'drink reading'	0.20
beber	*sustancia*	'drink substance'	11.93
beber	*alcohol*	'drink alcohol'	12.50
beber	*líquido*	'drink liquid'	22.33
tomar	*artrópodo*	'take arthropod'	0.20
tomar	*clase_alta*	'take high_class'	0.20
tomar	*conformidad*	'take conformity'	0.20
tomar	*postura*	'take posture'	49.83
tomar	*resolución*	'take resolution'	89.50
tomar	*control*	'take control'	114.75
tomar	*acción*	'take action'	190.18

Table 7 also shows that synsets located higher in WordNet hierarchy have higher values, as they accumulate the impact of the hyponym words that are below them (see for example *communication*, *liquid* or *action*). A simple ad-hoc strategy of weighting values in WordNet's hierarchy will be described also in Section 6.2.

Table 8. Selected combinations extracted from VCC

	verb	relation	noun	English gloss
1	*contar*	*con*	*permiso*	'to have permission'
2	*pintar*	<	*pintor*	'painter paints'
3	*golpear*	>	*balón*	'kick ball'
4	*solucionar*	>	*problema*	'solve problem'
5	*dar*	>	*señal*	'give signal'
6	*haber*	>	*incógnita*	'there is unknown quantity'
7	*poner*	*en*	*cacerola*	'put in pan'
8	*beber*	*de*	*fuente*	'drink from source'
9	*beber*	>	*vodka*	'drink vodka'

In the following sections we will show how we obtain information like that shown in Table 7, and then we will illustrate the usefulness of our method applying this information to word sense disambiguation (WSD).

6.1 Related Work

One of the first works on selectional preference extraction linked to WordNet senses was Resnik's [45]. It is devoted mainly to word sense disambiguation in English. Resnik assumed that a text annotated with word senses was a resource difficult to obtain, so he based his work on text tagged only morphologically. Subsequently, Agirre and Martinez [2, 3] worked linking verb usage with their arguments. In contrast with Resnik, Agirre and Martínez assumed the existence of a text annotated with word senses: Sem-Cor, in English. Other supervised WSD systems include JHU [57], which won the Senseval-2 competition, and a maximum entropy WSD system by Suarez and Palomar [51]. The first system combined, by means of a voting-based classifier, several WSD subsystems based on different methods: decision lists [56], cosine-based vector models, and Bayesian classifiers. The second system selected a best-feature selection for classifying word senses and a voting system. These systems had a score around 0.70 on the Senseval-2 tests.

We take into account that a resource such as Sem-Cor is currently not available for many languages (in particular, Spanish), and the cost of building it is high. Accordingly, we follow Resnik's approach, in the way of assuming that there is not enough quantity of text annotated with word senses. Furthermore, we consider that the WSD process must be completely automatic, so that all the text we use is automatically tagged with morphological and part-of-speech (POS) tags. Accordingly, our system is fully unsupervised.

Previous work on unsupervised systems has not achieved the same performance as with supervised systems: Carroll and McCarty [19] present a system that uses selectional preferences for WSD obtaining 69.1% precision and 20.5% recall; Agirre and Martinez [1] present another method, this time unsupervised. They use recall as the only performance measure, reporting 49.8%; Resnik [45] achieves 40% correct disambiguation.

In the next sections we describe our method and measure its performance.

6.2 Methodology

In order to obtain the selectional preferences linked to an ontology, we used the hypernym relations of Spanish EuroWordNet[8] 1.0.7 (S-EWN) as ontology, and the corpus described in [27] as a training corpus (VCC). This corpus of 38 million words is supposed to combine the benefits of a virtual corpus (e.g. the web as corpus), with those of a local corpus, see details in [27].

The text was morphologically tagged using the statistical tagger TnT by Thorsten Brants [9] trained with the corpus CLiC-TALP. This tagger has a performance of over 92%, as reported in [40].

After the text was tagged morphologically, several combinations were extracted for each sentence: (1) verb + noun to the left (subject), (2) verb + noun to the right (object), and (3) verb NEAR preposition + noun. Here, + denotes adjacency, while NEAR denotes co-occurrence within a sentence. Table 8 shows an example of the information obtained in this way. The symbol > means that the noun is to the right of the verb; the symbol < means that the noun appears to the left of the verb.

Once the combinations have been extracted, the noun for each combination was looked up in WordNet and an occurrence for the corresponding synset (with every sense) was recorded. Also the occurrence was recorded for each hyperonym of each its sense. A weighting factor was used so that words higher in the hierarchy (up to the root *entity*) have lower impact than the words in the lower part of the hierarchy. We used the weighting factor 1 / *level*. For example, for *drink vodka* found in the text, an occurrence of the combination *drink vodka* is recorded with the weight 1, also occurrences of *drink liquor* with the weight 0.5, *drink alcohol* with 0.33, etc. are recorded. For each combination, the weights of its occurrences are accumulated (summed up).

[8] S-EWN was Developed jointly by the University of Barcelona (UB), the Nacional University of Open Education (UNED), and the Polytechnic University of Catalonia (UPC), Spain.

atravesar canal: 'cross channel'
02342911n → **way** 3.00 → trough 8.83 → artifact 20.12 → unanimated_obect 37.10 → entity 37.63
02233055n → **conduit** 6.00 → way 3.00 → trough 8.83 → artifact 20.12 → unanimated_object 37.10 → entity 37.63
03623897n → **conduit** 5.00 → anatomic_structure 5.00 → body_part 8.90 → part 7.22 → entity 37.63
04143847n → **transmission** 1.67 → communication 3.95 → action 6.29
05680706n → **depression** 2.33 → geological_formation 2.83 → natural_object 14.50 → unanimated_object 37.10 → entity 37.63
05729203n → **water** 4.17 → unanimated_object 37.10 → entity 37.63

leer libro: 'read book'
01712031n → **stomach** 3.50 → internal_organ 3.00 → organ 3.08 → body_part 3.75 → part 4.35 → entity 41.51
02174965n → **product** 14.90 → creation 13.46 → artifact 34.19 → unanimated_object 36.87 → entity 41.51
04214018n → **section** 23.33 → writing 33.78 → written_language 25.40 → communication 55.28 → social_relation 43.86 → relation 42.38 → abstraction 44.18
04222100n → **publication** 16.58 → work 7.95 → product 14.90 → creation 13.46 → artifact 34.19 →unanimated_object 36.87 → entity 41.51
04545280n → **play** 4.50 → writing 33.78 → written_language 25.40 → communication 55.28 →social_relation 43.86 → relation 42.38 → abstraction 44.18

Fig. 5. Ontology with usage values for the combinations in Spanish *atravesar canal* 'cross channel' and *leer libro* 'read book'. Synsets labels were translated here from Spanish to English for the reader's convenience

Currently we have acquired 1.5 million of selectional preferences patterns linked to the WordNet synsets. Each pattern consists on a verb, a preposition (in some cases), and a synset. An example of the information obtained can bee seen in Figure 5. *Channel* has 6 senses listed by WordNet: *way*, *conduit*, *clear*, *conduit* (anatomic), *transmission*, *depression*, and *water*. The sense marked with the highest number of occurrences is *conduit*, while the one with fewer occurrences is *transmission*, in the sense of *channel of transmission* or *TV channel*, for example; one cannot *cross* a TV channel. Now consider *libro* 'book'; this Spanish word has five senses: *stomach*, *product*, *section*, *publication* and *work* / *play*. The first sense refers to the name in Spanish for an internal part of body. We can see that this is the sense with fewer occurrences (one cannot *read* an *organ*). The sense with the greatest number of occurrences is that related to *written_language*. This information can be used to disambiguate the sense of the word, given the verb with which it is used. In the next section we describe an experiment we ran to measure the performance of this method in the WSD task.

6.3 Experiments

Senseval is a series of competitions aimed to evaluation of word sense disambiguation programs, organized by the ACL-SIGLEX. The last competition took place in 2001 (the next one being scheduled for 2004). The data for this competition are available on-line. This competition included, among 10 languages, Spanish data, to which we applied our method. The evaluation set comprises slightly more than 1,000 sentences. Each sentence contains one word, for which the correct sense, among those listed for it in WordNet, is indicated.

Our evaluation showed that 577 of 931 cases were resolved (a recall of ~62%). Of those, 223 corresponded in a fine-grained way to the sense manually annotated (precision ca. 38.5%). These results are similar to those obtained by Resnik [45] for English, who obtained on average 42.55% for the relations verb—subject and verb—object only. Note that these results are much better than random selection of senses (around 28% as reported in [45]).

6.4 Discussion

Our results are lower than those of some other WSD systems. For example, Suarez and Palomar [51] report a score of 0.702 for noun disambiguation for the same evaluation set of Senseval-2. However, their system is supervised, whereas ours is unsupervised. In comparison with existing unsupervised WSD systems (i.e. [45, 19, 1]) our method

has a better recall, though lower precision in some cases. The latter is due the strategy of our method that considers only verb—noun relations, when sometimes the word sense is strongly linked to the preceding noun. This is particularly true for pairs of nouns that form a single prepositional phrase. For example, in the training text the following sentence appears: *La prevalecía del principio de libertad frente al principio de autoridad es la clave de Belle Epoque* 'The prevalence of the liberty principle in contrast with the authority principle is the key of Belle Epoque'. In this case, the sense of *autoridad* 'authority' is restricted more strongly by the preceding noun, *principio* 'principle', in contrast with the main verb: *es* 'is'. To determine the sense of *autoridad* by means of the combinations *is* < *authority* and *is of authority* is not the best strategy to disambiguate the sense of this word.

In order to improve our method, in the future we plan to include information on the usage of combinations of nouns.

7 Other Applications

Besides WSD, the information of selectional preferences obtained by this method can be used to solve important problems, such as syntactic disambiguation. For example, consider the phrase in Spanish *Pintó un pintor un cuadro*, lit. 'painted a painter a painting' meaning 'a painter painted a painting'. In Spanish it is possible to put the subject to the right of the verb. There is ambiguity, as it is not possible to decide which noun is the subject of the sentence. As Spanish is a language with rather free word order, even *Pintó un cuadro un pintor*, lit. 'painted a painting a painter' has the same meaning.

To decide which word is the subject (*painting* or *painter*) it is possible to consult the ontology linked with selectional preferences constructed with the method presented in this paper. First, we find statistically that the subject appears to the left of the verb in 72.6% of the times [37]. Then, searching for *un pintor* □*into* 'a painter painted' returns the following chain of hypernyms with occurrence values: *painter* → *artist* 1.00 → *creator* 0.67 → *human_being* 2.48 → *cause* 1.98. Finally, the search of *un cuadro* □*into* 'a painting painted' returns *scene*→ *situation* 0.42 → *state* 0.34. That is, *painter* (1.00) is more probable as subject than *painting* (0.42) for this sentence. A large-scale implementation of this method is a topic of our future work.

8 Conclusions

We have presented a simple and robust dependency parser for Spanish. It uses simple hand-made heuristic rules for the decisions on admissibility of structural elements and on word co-occurrence statistics for disambiguation. The statistics is learned from a large corpus, or obtained by querying an Internet search engine, in an unsupervised manner—i.e., no manually created tree-bank is used for training. In case if the parser cannot produce a complete parse tree, a partial structure is returned consisting of the dependency links it could recognize.

Comparison of the accuracy of our parser with two the available systems for Spanish we are aware of shows that our parser outperforms both of them.

Though a number of specific rules of the grammar are specific for Spanish, the approach itself is language-independent. As future work we plan to develop similar parsers for other languages, including English, for which the necessary preprocessing tools—such as POS tagger and lemmatizer—are available.

As other future work direction we could mention in the first place improvement of the system of grammar rules. The current rules sometimes do their job in a quick-and-dirty manner, which results in just the right thing to do in most of the cases, but can be done with greater attention to details.

We presented a method to extract selectional preferences of verbs linked to an ontology. It is useful to solve natural language text processing problems that require information about the usage of words with a particular verb in a sentence. Specifically, we presented an experiment that applies this method to disambiguate word senses. The results of this experiment show that there is still a long way to improve unsupervised WSD methods using selectional preferences; however, he have identified specific points to improve our method under the same line of pattern-based statistical methods combined with knowledge.

The work presented here resulted in the following main contributions:

- DILUCT: A syntactic dependency analyzer for Spanish; we made test against similar analyzers, yielding a better performance.
- A selectional preferences database for Spanish. It contains 3 million of different combinations; 0.43 million include prepositions.
- Several algorithms for PP attachment disambiguation. We improved existing algorithms.
- Creation of a distributional thesaurus for Spanish following Lin's method.
- Comparison of manual dictionaries against automatically obtained dictionaries. The result of this research suggest that dictionaries obtained automatically using a computer can substitute dictionaries created manually in certain tasks, saving years of human work.
- A method to convert automatically an annotated corpus from constituents to dependencies.

In the future, we plan to evaluate the usefulness of our parser in real-world tasks of information retrieval, text mining, and constructing semantic representation of the text, such as conceptual graphs.

References

1. **Agirre E., D. Martínez.** Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* EMNLP, Barcelona, Spain, 2004.
2. **Agirre, E. D. Martinez.** Learning class-to-class selectional preferences. In: *Proceedings of the Workshop Computational Natural Language Learning* (CoNLL-2001), Toulousse, France, 6-7 july, 2001.
3. **Agirre, E., D. Martinez.** Integrating selectional preferences in WordNet. In: *Proceedings of the first International WordNet Conference*, Mysore, India, 21-25 January, 2002.
4. **Apresyan, Yuri D., Igor Boguslavski, Leonid Iomdin, Alexandr Lazurski, Nikolaj Pertsov, Vladimir Sannikov, Leonid Tsinman.** *Linguistic Support of the ETAP-2 System* (in Russian). Moscow, Nauka, 1989.
5. **Bolshakov, Igor A.** A Method of Linguistic Steganography Based on Collocationally-Verified Synonymy. *Information Hiding 2004*, *Lecture Notes in Computer Science*, 3200 Springer-Verlag, 2004, pp. 180–191.
6. **Bolshakov, Igor A., Alexander Gelbukh.** Lexical functions in Spanish. Proc. *CIC-98, Simposium Internacional de Computación*, Mexico, pp. 383–395; www.gelbukh.com/CV/Publications/1998/ CIC-98-Lexical-Functions.htm, 1998.
7. **Bolshakov, Igor A., Alexander Gelbukh.** A Very Large Database of Collocations and Semantic Links. Proc. NLDB-2000: 5th Intern. Conf. on Applications of Natural Language to Information Systems, France, *Lecture Notes in Computer Science* N 1959, Springer-Verlag, 2000, pp. 103–114.
8. **Bolshakov, Igor A., Alexander Gelbukh.** On Detection of Malapropisms by Multistage Collocation Testing. *NLDB-2003, 8th Int. Conf. on Application of Natural Language to Information Systems*. Bonner Köllen Verlag, 2003, pp. 28–41.
9. **Brants, T.,** TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA, 2000.
10. **Brants, Thorsten.** TNT–A Statistical Part-of-Speech Tagger. In: Proc. *ANLP-2000, 6th Applied NLP Conference*, Seattle, 2000.
11. **Brill, Eric, Philip Resnik.** A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation, In *Proceedings of COLING-1994*, 1994.
12. **Briscoe, Ted. John Carroll, Jonathan Graham and Ann Copestake.** Relational evaluation schemes. In: *Procs. of the Beyond PARSEVAL Workshop, 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Gran Canaria, 2002, 4–8.
13. **Calvo, Hiram, Alexander Gelbukh.** Extracting Semantic Categories of Nouns for Syntactic Disambiguation from Human-Oriented Explanatory Dictionaries, In *Computational Linguistics and Intelligent Text Processing*, Springer LNCS 2945, 2004.

14. **Calvo, Hiram, Alexander Gelbukh.** Improving Prepositional Phrase Attachment Disambiguation Using the Web as Corpus, In A. Sanfeliu and J. Shulcloper (Eds.) *Progress in Pattern Recognition*, Springer LNCS 2905, 2003, pp. 604-610
15. **Calvo, Hiram, Alexander Gelbukh.** Natural Language Interface Framework for Spatial Object Composition Systems. *Procesamiento de Lenguaje Natural* 31, 2003.
16. **Calvo, Hiram, Alexander Gelbukh.** Acquiring Selectional Preferences from Untagged Text for Prepositional Phrase Attachment Disambiguation. In: *Proc. NLDB-2004*, *Lecture Notes in Computer Science* 3136, 2004, pp. 207–216.
17. **Calvo, Hiram. Alexander Gelbukh, Adam Kilgarriff.** Distributional Thesaurus versus WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. In: *Computational Linguistics and Intelligent Text Processing* (CICLing-2005). *LNCS* 3406, Springer-Verlag, 2005, pp. 177–188.
18. **Carreras, Xavier, Isaac Chao, Lluis Padró, Muntsa Padró. FreeLing:** An Open-Source Suite of Language Analyzers. *Proc. 4th Intern. Conf. on Language Resources and Evaluation (LREC-04)*, 2004, Portugal.
19. **Carroll, J., D. McCarthy.** Word sense disambiguation using automatically acquired verbal preferences. In *Computers and the Humanities*, 34(1-2), Netherlands, 2000.
20. **Chomsky, Noam.** *Syntactic Structures*. The Hague: Mouton & Co, 1957.
21. **Civit, Montserrat, and Maria Antònia Martí.** Estándares de anotación morfosintáctica para el español. *Workshop of tools and resources for Spanish and Portuguese*. IBERAMIA 04, Mexico, 2004.
22. **Copestake, Ann, Dan Flickinger, Ivan A. Sag.** *Minimal Recursion Semantics. An introduction.* CSLI, Stanford University, 1997.
23. **Debusmann, Ralph, Denys Duchier, Geert-Jan M. Kruijff,** Extensible Dependency Grammar: A New Methodology. In: *Recent Advances in Dependency Grammar. Proc. of a workshop at COLING-04*, Geneve, 2004
24. **Díaz, Isabel, Lidia Moreno, Inmaculada Fuentes, Oscar Pastor.** Integrating Natural Language Techniques in OO-Method. In: Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing* (CICLing-2005), *Lecture Notes in Computer Science* 3406, Springer-Verlag, 2005, pp. 560–571.
25. **Dik, Simon C.**, *The Theory of Functional Grammar. Part I: The structure of the clause*. Dordrecht, Foris, 1989.
26. **Dirk, Lüdtke, Satoshi Sato.** Fast Base NP Chunking with Decision Trees - Experiments on Different POS Tag Settings. In Gelbukh, A. (ed) *Computational Linguistics and Intelligent Text Processing*, Springer LNCS, 2003, pp. 136-147.
27. **Gelbukh, A., G. Sidorov, L. Chanona.** Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet. In: Julio Gonzalo, Anselmo Peñas, Antonio Ferrández, eds.: *Proc. Multilingual Information Access and Natural Language Processing, International Workshop,* in IBERAMIA-2002, VII Iberoamerican Conference on Artificial Intelligence, Seville, Spain, November 12-15, 2002, 7–14.
28. **Gelbukh, A., S. Torres, H. Calvo.** Transforming a Constituency Treebank into a Dependency Treebank. Submitted to *Procesamiento del Lenguaje Natural* No. 34, Spain, 2005.
29. **Gelbukh, Alexander, Grigori Sidorov, Francisco Velásquez.** Análisis morfológico automático del español a través de generación. *Escritos*, N 28, 2003, pp. 9–26.
30. **Gladki, A. V.** *Syntax Structures of Natural Language in Automated Dialogue Systems* (in Russian). Moscow, Nauka, 1985.
31. **Kudo, T., Y. Matsumoto.** Use of Support Vector Learning for Chunk Identification. In *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal, 2000.
32. **Lara, Luis Fernando.** *Diccionario del español usual en México*. Digital edition. Colegio de México, Center of Linguistic and Literary Studies, 1996.
33. **Mel'čuk, Igor A.** Meaning-text models: a recent trend in Soviet linguistics. *Annual Review of Anthropology* 10, 1981, 27–62.
34. **Mel'čuk, Igor A.** Dependency Syntax: Theory and Practice. State U. Press of NY, 1988.

35. **Mel'čuk, Igor A.** Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In: L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: Benjamins, 1996, 37–102.
36. **Miller, G.** WordNet: An on-line lexical database, In *International Journal of Lexicography*, 3(4), December 1990, pp. 235–312.
37. **Monedero, J., González, J. Goñi, C. Iglesias, A. Nieto.** Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS. In *Actas del XI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* SEPLN 95, Bilbao, Spain, 1995, 241—254.
38. **Montes-y-Gómez, Manuel, Alexander F. Gelbukh, Aurelio López-López.** Text Mining at Detail Level Using Conceptual Graphs. In: Uta Priss *et al*. (Eds.): *Conceptual Structures: Integration and Interfaces*, 10th Intern. Conf. on Conceptual Structures, ICCS-2002, Bulgaria. *LNCS* 2393, Springer-Verlag, 2002, pp. 122–136.
39. **Montes-y-Gómez, Manuel, Aurelio López-López, and Alexander Gelbukh.** Information Retrieval with Conceptual Graph Matching. *Proc. DEXA-2000, 11th Intern. Conf. DEXA, England, LNCS* 1873, Springer-Verlag, 2000, pp. 312–321.
40. **Morales-Carrasco, R., A. Gelbukh.** Evaluation of TnT Tagger for Spanish. In *Proc. Fourth Mexican International Conference on Computer Science*, Tlaxcala, Mexico, September 8-12, 2003.
41. **Pollard, Carl, and Ivan Sag.** *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, IL and London, UK, 1994.
42. **Prescher, Detlef, Stefan Riezler, and Mats Rooth.** Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarland University, Saarbrücken, Germany, 2000.
43. **Ratnaparkhi, Adwait, Jeff Reynar, and Salim Roukos.** A Maximum Entropy Model for Prepositional Phrase Attachment. In *Proceedings of the ARPA Human Language Technology Workshop*, 1994, pp. 250-255.
44. **Resnik, P.** Selectional Constraints: An Information-Theoretic Model and its Computational Realization, *Cognition*, 61, November, 1996, 127–159.
45. **Resnik, P.** Selectional preference and sense disambiguation, *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C., USA, April 4-5, 1997.
46. **Resnik, P.** *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. Thesis, University of Pennsylvania, December, 1993.
47. **Sag, Ivan, Tom Wasow, and Emily M. Bender.** *Syntactic Theory. A Formal Introduction* (2nd Edition). CSLI Publications, Stanford, CA, 2003
48. **Sebastián, N., M. A. Martí, M. F. Carreiras, and F. Cuestos.** *LEXESP, léxico informatizado del español*, Edicions de la Universitat de Barcelona, 2000.
49. **Sowa, John F. 1984.** *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley Publishing Co., Reading, MA, 1984.
50. **Steele, James (ed.).** *Meaning-Text Theory. Linguistics, Lexicography, and Implications*. Ottawa: Univ. of Ottawa Press, 1990.
51. **Suárez, A., M. Palomar.** A Maximum Entropy-based Word Sense Disambiguation System. In: Hsin-Hsi Chen and Chin-Yew Lin, eds.: *Proceedings of the 19th International Conference on Computational Linguistics*, COLING 2002, Taipei, Taiwan, vol. 2, 2002, 960—966.
52. **Tapanainen, Pasi.** *Parsing in two frameworks: finite-state and functional dependency grammar*. Academic Dissertation. University of Helsinki, Language Technology, Department of General Linguistics, Faculty of Arts, 1999.
53. **Tesnière, Lucien.** *Eléments de syntaxe structurale.* Paris: Librairie Klincksieck, 1959.
54. Volk, Martin. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceeding of Corpus Linguistics* 2001. Lancaster, 2001.
55. **Weinreich, Uriel.** *Explorations in Semantic Theory*, Mouton, The Hague, 1972.

56. **Yarowsky, D.,** Hierarchical decision lists for word sense disambiguation. In *Computers and the Humanities*, 34(2), 2000, 179–186.
57. **Yarowsky, D.,** S. Cucerzan, R. Florian, C. Schafer, R. Wicentowski. The Johns Hopkins SENSEVAL-2 System Description. In: Preiss and Yarowsky, eds.: *The Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, 2001, 163–166.
58. **Yuret, Deniz.** *Discovery of Linguistic Relations Using Lexical Attraction*, PhD thesis, MIT, 1998.

Hiram Calvo *was born in Mexico in 1978. He obtained his Master degree in Computer Science in 2002 from National Autonomous University of Mexico (UNAM), with a thesis on mathematical modeling, and his Ph. D. degree in Computer Science (with honors) in 2006 from the Computing Research Center (CIC) of the National Polytechnic Institute (IPN), Mexico, with a thesis on natural language processing. Since 2006 he is a lecturer at the Computing Research Center (CIC) of the National Polytechnic Institute (IPN). He was awarded with the Lázaro Cárdenas Prize (2006) as the best Ph. D. student of IPN in the area of physics and mathematics (this Prize is handed personally by the President of Mexican Republic) and a prize for the best Ph. D. thesis of IPN in the area. He is author of more than 15 scientific publications.*

Alexander Gelbukh. *He was botn in Moscow, Russia, in 1962. He obtained his Master degree in Mathematics in 1990 from the department of Mechanics and Mathematics of the "Lomonosov" Moscow State University, Russia, and his Ph. D. degree in Computer Science in 1995 from the All-Russian Institute of the Scientific and Technical Information (VINITI), Russia. Since 1997 he is Professor and the head of the Natural Language and Text Procesing Laboratory of the Computing Research Center (CIC), National Polytechnic Institute (IPN), Mexico City. He is an academician of Mexican Academy of Sciences since 2000 and National Researcher of Mexico (SNI) since 1998; author of more than 300 publications on computational linguistics; see www.Gelbukh.com.*