

# Meaning Representation for Automatic Extraction of Lexical Functions

Olga Kolesnikova

Centro de Investigación en Computación,  
Instituto Politécnico Nacional,  
Mexico, D.F. 07738, Mexico.  
kolesolga@gmail.com

**Resumen.** Lexical functions formalize semantic and syntactic relations between lexical units, given that meaning of an individual word largely depends on various relations connecting it to other words in context. Collocational relation is a type of institutionalized lexical relations that holds between the base and its partner in a collocation in contrast to free word combination where both words are used in their typical meaning. Collocation are important for natural language processing because collocation comprises the restrictions on how words can be used together. The formalism of lexical functions is a means of representing such information. If collocations are annotated with lexical functions in a computer readable dictionary, it allows effective use of collocations in natural language applications including parsers, high quality machine translation, periphrasis systems and computer-aided learning of lexica. In order to create such applications, we need to extract lexical functions from corpora automatically. For this, we represent the lexical meaning of a given word with a set of all its hypernyms extracted from the Spanish WordNet.

**Keywords:** natural language processing, lexical functions, semantic representation, machine learning.

## 1 Introducción

Lexical function is a concept that formalizes semantic and syntactic relations between lexical units. Relations between words are a vital part of any natural language system. Meaning of an individual word largely depends on various relations connecting it to other words in context. In particular, collocational relation is a type of institutionalized lexical relations that holds between the base and its partner in a collocation.

A collocation typically consists on the main word, or base, and the word that collocates with it, or collocate. Examples of collocations are *give a lecture*, *make a decision*, *lend support*, where the bases are *lecture*, *decision*, *support* and the partners, termed collocates, are *give*, *make*, *lend*. Collocations are opposed to free word combination where both words are used in their typical meaning, such as, for example, *give a book*, *make a dress*, *lend money*.

Knowledge of collocation is important for natural language processing and its applications because collocation comprises the restrictions on how words can be used together. There are many methods to extract collocations automatically but their result is a plain list of collocations. Such lists are more valuable if collocations are tagged with semantic and grammatical information.

The formalism of lexical functions is a means of representing such information. If collocations are annotated with lexical functions in a computer readable dictionary, it will allow effective use of collocations in natural language applications including parsers [Gelbukh and Sidorov 2006, Bolshakov and Gelbukh 2004], high quality machine translation [Bolshakov and Gelbukh 2001], periphrasis system and computer-aided learning of lexica [Bolshakov and Gelbukh 2002].

In order to create such applications, we need to extract lexical functions from corpora automatically. It is our intent to extract Spanish verb-noun collocations belonging to a given lexical function from corpora. To achieve this, we represent the lexical meaning of a given word with a set of all its hypernyms. This allows us to use machine-learning techniques for predicting lexical functions as values of the class variable for unseen collocations. We extract such hypernyms from the Spanish WordNet. Our experiments show that machine learning is feasible to achieve the task of automatic detection of lexical functions.

Relatively little research has been done so far on automatic detection of lexical functions. In fact, there are only two papers that report results on performance of a few machine-learning algorithms on classifying collocations according to the typology of lexical functions [Wanner 2004, Wanner *et al.* 2006].

In this paper, we first review various definitions of collocations given in existing literature, to clarify thoroughly the goal of our research. Then we summarize existing work on collocation extraction, in particular on extraction of lexical functions: we consider the work done in [Wanner 2004], [Wanner *et al.* 2006] and comment on another research on automatic extraction of lexical functions [Alonso Ramos *et al.* 2008] based on an approach different from the work in [Wanner 2004] and [Wanner *et al.* 2006]. We discuss the three statements, or hypotheses, made in [Wanner *et al.* 2006]. Next, we describe the data used in our experiments. Finally, we present our methodology for meaning representation for collocation extraction.

## 1.1 Definitions of Collocation

Our goal is to extract lexical functions, which are a particular type of a collocation. There is no consensus on what a collocation is. Here we review numerous different approaches to the definition of a collocation.

With each definition, additional information is given as to the source of definition, the criterion used to distinguish collocations from free word combinations, and some our comments on the particular definition.

[Firth 1957]: *Collocations of a given word are statements of the habitual or customary places of that word.* Lexical criterion: a word is used in a fixed position with respect to a given word; statistical criterion: frequency of word co-occurrence.

[Firth 1957] first introduced the term ‘collocation’ from Latin *collocatio* which means ‘bringing together, grouping’. Firth believes that speakers make ‘typical’ common lexical choices in collocational combinations. Collocation is a concept in Firth’s theory of meaning: “Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*.”

[Halliday 1961]: *Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c ...* Lexical criterion: a word is used a fixed position with respect to a given word. Statistical criterion: high co-occurrence frequency.

If a lexical item is used in the text, then its collocate has the highest probability of occurrence at some distance from the lexical item. Collocations cut across grammar boundaries: e.g., *he argued strongly* and *the strength of his argument* are grammatical transformations of the initial collocation *strong argument*.

[Hausmann 1984]: *Collocations are binary word-combinations, consist of words with limited combinatorial capacity, they are semi-finished products of language, affine combinations of striking habitualness. In a collocation one partner determines, another is determined. In other words: collocations have a basis and a co-occurring collocate.* Lexical criterion: the lexical choice of the collocate depends on the basis.

Word combinations are classified word-combinations according to the features fixed vs. non-fixed, and in this classification collocations are belong to the category of non-fixed affine combinations. Internal structure of collocation: collocation components have functions of a basis and a collocate, and the basis (not the speaker) ‘decides’ what the collocate will be.

[Benson 1986]: *Collocation is a group of words that occurs repeatedly, i. e. recurs, in a language. Recurrent phrases can be divided into grammatical collocations and lexical collocations. Grammatical collocations consist of a dominant element and a preposition or a grammatical construction: fond of, (we reached) an agreement that... Lexical collocations do not have a dominant word, their components are "equal": to come to an agreement, affect deeply, weak tea.* Functional criterion: collocations are classified according to function of collocational elements. Statistical criterion: high co-occurrence frequency.

This is a broad understanding of collocation. Classification of collocations according to their compositional structure is given.

[Benson 1990]: *Collocations should be defined not just as ‘recurrent word combinations’, <but as> ‘ARBITRARY recurrent word combinations’.* Lexical criterion: arbitrariness and recurrency.

‘Arbitrary’ here is opposed to ‘regular’ means that collocations are not predictable and cannot be translated word by word.

[Van Roey 1990]: *Collocation is “that linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its ‘synonyms’ because of*

*constraints which are not on the level of syntax or conceptual meaning but on that of usage.*” Statistical criterion: high co-occurrence frequency in corpora.

Van Roey summarizes statistical view stated by Halliday in terms of expression or ‘usage’. A collocate can thus simply be seen as any word which co-occurs within an arbitrary determined distance or *span* of a central word or *node* at the frequency level at which the researcher can say that the co-occurrence is not accidental. This approach is also textual in that it relies solely on the ability of the computer program to analyze large amounts of computer-readable texts.

[Cowie 1994]: *Collocations are associations of two or more lexemes (or roots) recognized in and defined by their occurrence in a specific range of grammatical constructions.* Structural criterion: collocations are distinguished by patterns.

Collocations are classified by Cowie into types according to their grammatical patterns.

[Howarth 1996]: *In his lexical continuum model, collocations as composite units are placed on a sliding scale of meaning and form from relatively unrestricted (collocations) to highly fixed (idioms). Restrictive collocations are fully institutionalised phrases, memorized as wholes and used as conventional form-meaning pairings.* Syntactic criterion: commutability – the extent to which the elements in the expression can be replaced or moved (make/reach/take decision vs. shrug one’s shoulders). Semantic criterion: motivation – the extent to which the semantic origin of the expression is identifiable (move the goalposts = to change conditions for success vs. shoot the breeze = to chatter, which is an opaque idiom).

Classification includes four types of expressions with no reference to frequency of occurrence:

- free collocation: *blow a trumpet* = to play a trumpet,
- restrictive collocation: *blow a fuse* = to destroy a fuse/to get angry,
- figurative idiom: *blow your own trumpet* = to sell oneself excessively,
- pure idiom: *blow the gaff* = to reveal a concealed truth.

The problem with this classification is that is difficult to determine what is meant by ‘syntactically fixed’, ‘unmotivated’ or ‘opaque’. This is seen in the ambiguous example of *to blow a fuse*.

[Sinclair et al. 2004]: *Collocation is the co-occurrence of two items in a text within a specified environment. Significant collocation is regular collocation between two items, such that they co-occur more often than their respective frequencies. Casual collocations are “non-significant” collocations.* Lexical criterion: recurrency of co-occurrence. Statistical criterion: high co-occurrence frequency.

The degree of significance for an association between items is here determined by such statistic tests as Fischer’s Exact Test or Poisson Test.

[Mel’čuk 1996]: *Collocation is a combination of two lexical items in which the semantics of one of the lexical items (the base) is autonomous from the combination it appears in, and where the other lexical item (the collocate) adds semantic features to the semantics of the base.* [Gledhill 2000] explains that for Mel’čuk a collocation is a semantic function operating between two or more words in which one of the words keeps its

'normal' meaning. Semantic criterion: the meaning of a collocation is not inferred from the meaning of the base combined with meaning of the collocate.

According to Mel'čuk, semantics of a collocation is not the meaning of the base + the meaning of the collocate, but rather the meaning of the base + some additional meaning that are included in the meaning of the base. In particular: '...the concept of collocation is independent of grammatical categories: the relationship, which holds between the verb *argue* and the adverb *strongly* is the same as that holding between the noun *argument* and the adjective *strong*' [Fontenelle 1994].

## 2 Related Work

Bolshakov and Gelbukh [1998] studied lexical functions in Spanish on a number of examples. They described various types of such collocations and discussed how some functions can be combined to give rise to new meanings. They also studied classification of collocations according to the meaning of the words being combined [Bolshakov and Gelbukh, 2000].

In 2004, L. Wanner proposed to view the task of LF detection as automatic classification of collocations according to LF typology. To fulfill this task, the nearest neighbor machine learning technique was used. Datasets included Spanish verb-noun pairs annotated with nine LFs: CausFunc<sub>0</sub>, Caus<sub>2</sub>Func<sub>1</sub>, IncepFunc<sub>1</sub>, FinFunc<sub>0</sub>, Oper<sub>1</sub>, ContOper<sub>1</sub>, Oper<sub>2</sub>, Real<sub>1</sub>, Real<sub>2</sub>. Verb-noun pairs were divided in two groups. In the first group, nouns belonged to the semantic field of emotions; in the second groups nouns were field-independent. As a source of information for building the training and test sets, hypernymy hierarchy of the Spanish part of EuroWordNet was used.

The words in the training set were represented by their hypernyms, Basic Concepts and Top Concepts. The average *F-measure* of about 70% was achieved in these experiments. The best result for field-independent nouns was F-measure of 76.58 for CausFunc<sub>0</sub> with the meaning 'cause the existence of the situation, state, etc.' The Causer is the subject of utterances with CausFunc<sub>0</sub>.

In [Wanner *et al.* 2006], four machine learning methods were applied to classify Spanish verb-noun collocations according to LFs, namely Nearest Neighbor technique, Naïve Bayesian network, Tree-Augmented Network Classification technique and a decision tree classification technique based on the ID3-algorithm. As in [Wanner 2004], experiments were carried out for two groups of verb-noun collocations: nouns of the first group belonged to the semantic field of emotions; nouns of the second group were field-independent. Lexical functions were also identical with [Wanner 2004] as well as data representation. The best results for field-independent nouns were shown by ID3 algorithm (F-measure of 0.76) for Caus<sub>2</sub>Func<sub>1</sub> with the meaning 'cause something to be experienced / carried out / performed', and by the Nearest Neighbor technique (F-measure of 0.74) for Oper<sub>1</sub> with the meaning 'perform / experience / carry out something'. The Causer is the subject of utterances with Caus<sub>2</sub>Func<sub>1</sub>, and the Agent is the direct object of the verb which is the value of Caus<sub>2</sub>Func<sub>1</sub>. In utterances with Oper<sub>1</sub>, the Agent is the subject.

As we are interested in experiments with verb-noun collocations where the nouns have various semantics, i.e., the nouns are field-independent, Tables 1–4 summarizes the results for field-independent nouns only in [Wanner 2004] and [Wanner *et al.* 2006].

Table 1 gives the meaning of lexical functions used in experiments only with field-independent nouns [Wanner 2004]. We give examples in Spanish with literal translation in English. After the name of a lexical function, we give three figures with the following meaning:

- the number of examples of a given LF in the training set;
- the number of examples of a given LF in the test set;
- the total number of examples of a given LF in the training set and in the test set.

**Table 1.** Data in [Wanner 2004]

Name	Meaning	Examples in Spanish	Lit. translation in English
Oper <sub>1</sub> 35 + 15 = 50	<i>experience, perform, carry out something</i>	<i>dar golpe presentar una demanda hacer campaña dictar la sentencia</i>	<i>give a blow present a demand make a campaign dictate a sentence</i>
Oper <sub>2</sub> 33 + 15 = 48	<i>undergo, be source of</i>	<i>someterse a un análisis afrentar un desafío hacer examen tener la culpa</i>	<i>submit oneself to analysis face a challenge make exam have guilt</i>
CausFunc <sub>0</sub> 38 + 15 = 53	<i>cause the existence of the situation, state, etc.</i>	<i>dar la alarma celebrar elecciones provocar una crisis publicar una revista</i>	<i>give the alarm celebrate elections provoke a crisis publish a magazine</i>
Real <sub>1</sub> 37 + 15 = 52	<i>act accordingly to the situation, use as foreseen</i>	<i>ejercer la autoridad utilizar el teléfono hablar la lengua cumplir la promesa</i>	<i>exercise authority use a telephone speak a language keep a promise</i>
Real <sub>2</sub> 38 + 15 = 53	<i>react accordingly to the situation</i>	<i>responder a objeción satisfacer un requisito atender la solicitud rendirse a persuasión</i>	<i>respond to an objection satisfy a requirement attend an application surrender to persuasion</i>

Table 2 lists LFs with respective number of examples in [Wanner *et al.* 2006] for verb-noun combinations with field-independent nouns.

**Table 2.** Data in [Wanner *et al.* 2006]

LF	Number of Examples
CausFunc <sub>0</sub>	53
Oper <sub>1</sub>	87
Oper <sub>2</sub>	48
Real <sub>1</sub>	52
Real <sub>2</sub>	53

Table 3 presents the results reported in the referenced paper, in terms of accuracy by each lexical function.

**Table 3.** Results in [Wanner 2004]

F-measure/LF	CausFunc <sub>0</sub>	Oper <sub>1</sub>	Oper <sub>2</sub>	Real <sub>1</sub>	Real <sub>2</sub>
field-independent nouns	76.58	60.93	75.85	74.06	58.32

Finally, Table 4 shows the results in [Wanner *et al.* 2006]; the values of precision, recall and F-measure are given in the following format: <precision> | <recall> | <F-measure>. Not all four machine-learning methods in Table 4 were applied to all LFs; if experiments were not made for a particular method and LF, N/A is put instead of precision, recall, and F-measure.

**Table 4.** Results in [Wanner *et al.* 2006]

LF	Machine learning technique			
	NN	NB	TAN	ID3
CausFunc <sub>0</sub>	0.59   0.79   0.68	0.44   0.89   0.59	0.45   0.57   0.50	N/A
Caus <sub>2</sub> Func <sub>1</sub>	N/A	N/A	N/A	0.53   0.65   0.50
FinFunc <sub>0</sub>	N/A	N/A	N/A	0.53   0.40   0.40
IncepFunc <sub>1</sub>	N/A	N/A	N/A	0.40   0.48   0.40
Oper <sub>1</sub>	0.65   0.55   0.60	0.87   0.64   0.74	0.75   0.49   0.59	0.52   0.51   0.50
Oper <sub>2</sub>	0.62   0.71   0.66	0.55   0.21   0.30	0.55   0.56   0.55	N/A
ContOper <sub>1</sub>	N/A	N/A	N/A	0.84   0.57   0.68
Real <sub>1</sub>	0.58   0.44   0.50	0.58   0.37   0.45	0.78   0.36   0.49	N/A
Real <sub>2</sub>	0.56   0.55   0.55	0.73   0.35   0.47	0.34   0.67   0.45	N/A

On the other hand, [Alonso Ramos *et al.* 2008] proposed an algorithm for extracting collocations following the pattern “support verb + object” from FrameNet corpus of examples [Ruppenhofer *et al.* 2006] and checking if they are of the type Oper<sub>n</sub>. This work takes advantage of syntactic, semantic and collocation annotations in the FrameNet corpus, since some annotations can serve as indicators of a particular LF. The authors tested the proposed algorithm on a set of 208 instances. The algorithm showed accuracy of 76%. The researchers conclude that extraction and semantic classification of collocations is feasible with semantically annotated corpora. This statement sounds logical because the formalism of lexical function captures the correspondence between the semantic valence of the keyword and the syntactic structure of utterances where the keyword is used in a collocation together with the value of the respective LF.

## 2.1 Hypothesis Stated by Wanner *et al.*

Wanner *et al.* [2006] experimented with the same type of lexical data as in [Wanner 2004], i.e. verb-noun pairs. The task is to answer the question: what kind of collocational features are fundamental for human distinguishing among collocational types. The authors view collocational types as LFs, i.e. a particular LF represents a certain type of collocations.

Three hypotheses are put forward as possible solutions, and to model every solution, an appropriate machine learning technique is selected. Below we list the three hypotheses and the selected machine learning techniques.

1. Collocations can be recognized by their similarity to the prototypical sample of each collocational type; this strategy is modeled by the Nearest Neighbor technique.
2. Collocations can be recognized by similarity of semantic features of their elements (i.e., base and collocate) to semantic features of elements of the collocations known to belong to a specific LF; this method is modeled by Naïve Bayesian network and a decision tree classification technique based on the ID3-algorithm.
3. Collocations can be recognized by correlation between semantic features of collocational elements; this approach is modeled by Tree-Augmented Network Classification technique.

It should be mentioned, that having proposed three hypotheses, the authors have not yet demonstrated their validity by comparing the performance of many machine-learning techniques known today, but applied only four learning algorithms to illustrate that three human strategies mentioned above are practical.

## 2.2 Automatic Detection of Semantic Relations

There has been some research done on semantic relations in word combinations, for example, one that deals with automatic assignment of semantic relations to English noun-modifier pairs in [Nastase and Szpakowicz 2003, Nastase *et al.* 2006]. Though in our work, verb-noun combinations are treated, we believe that the principles of choosing data representation and machine learning techniques for detection of semantic relations between a noun and a modifier can be used to detect semantic relations in verb-noun pairs.

The underlying idea is the same: learning the meaning of word combinations. In [Nastase and Szpakowicz 2003, Nastase *et al.* 2006], the researchers examined the following relations: causal, temporal, spatial, conjunctive, participant, and quality. They used two different data representations: the first is based on WordNet relations, the second, on contextual information extracted from corpora. They applied memory-based learning, decision tree induction and Support Vector Machine. The highest F-measure of 0.847 was achieved by C5.0 decision tree to detect temporal relation based on WordNet representation.

## 3 Data Sets Used in Our Experiments

We have created a unique lexical resource of Spanish lexical functions in order to compile training sets for machine learning experiments.

### 3.1 Data for the Training Sets

For training and for testing, we used in our experiments the data sets presented in the sequel.



### **Lexical Resources**

Lexical resources are widely used in natural language processing and their role is difficult to overestimate. Lexical resources vary significantly in language coverage and linguistic information they include, and have many forms: word lists, dictionaries, thesauri, ontologies, glossaries, concordances, etc.

For Spanish, this diversity of forms can be illustrated with the following lexicographic works:

- A Medieval Spanish Word List [Oelschläger 1940],
- Diccionario de la Lengua Española (Dictionary of the Spanish Language) [RAE 2001],
- Streetwise Spanish Dictionary/Thesaurus [McVey and Wegmann 2001],
- Spanish part of EuroWordNet [Vossen 1998], an electronic lexical ontology,
- Glosario de voces comentadas en ediciones de textos clásicos [Fontecha 1941],
- Concordancia electrónica de la Biblia online (for Reina Valera version, 1960) [CEB].

Machine-readable resources are of special interest, since they comprise an integral part of computer systems aimed at automatic language treatment and language generation.

Though computerized lexicography has achieved a significant progress over last years, compilation of high quality dictionaries still requires a lot of manual work. In such a multi-faceted area as computational linguistics, it is difficult sometimes to find an adequate lexical resource (and for the language you need) for a specific research task or application.

One way to solve this problem is to develop computational procedures that can adjust existing resources to the demands of a researcher. However, this is not always effective. Certainly, the best solution of this problem is to compile a new lexical resource, but this is not always feasible in view of its cost.

We present a list of most frequent Spanish verb-noun pairs which contains semantically annotated collocations and free word combinations. It is a machine readable lexical resource where each verb-noun pair is associated with the following linguistic data:

1. whether a pair is a free word combination or a collocation;
2. if a verb-noun pair is a collocation, it is marked with lexical functions;
3. word senses of the Spanish WordNet [Vossen 1998, SpWN] are assigned to both elements of the verb-noun pair.

### **Existing Lexical Resources**

A number of lexical resources contain lexical functions. Almost all of them are not specialized dictionaries of lexical functions, but include lexical functions together with other linguistic information.

The concept of lexical function was originally proposed by researchers of the Russian semantic school. Lexical functions have been applied there for description of lexica and machine translation. A dictionary in Russian compiled by Apresjan [referenced in

Apresjan 2004] for the machine translation system ETAP includes more than 100 lexical functions with definitions and examples. For instance, for the verbal lexical function Oper<sub>1</sub>, the dictionary contains several hundreds of samples.

Lexical functions are used to describe the word's combinatory power in Explanatory Combinatorial Dictionaries compiled for Russian [Mel'čuk and Zholkovskij 1984] and for French [Mel'čuk *et al.* 1984, 1988]. For every word, its lexical entry includes a list of lexical functions applicable to it with their respective values. For French, an on-line dictionary, the DiCo, is referenced in [Wanner 2004] but we could not access it on the web.

For Spanish, there exists a dictionary of collocations, *Diccionario de colocaciones del Español* [DiCE] [Alonso Ramos 2003] annotated with lexical functions, but the DiCE is limited only to nouns belonging to the semantic field of emotions. [Sanromán 1998, 2003] compiled collections of Spanish collocations also for emotion nouns classified in terms of lexical functions. [Wanner 2004, Wanner *et al.* 2006] used Sanromán's collections for machine learning experiments, and for the same purpose, compiled additional lists of Spanish verb-noun collocations annotated with lexical functions. In the additional lists nouns were semantically field independent. The overall number of LF instances in the latter lists were 256 [Wanner 2004] and 293 [Wanner *et al.* 2006]. Unfortunately, these lists are no longer available in full.

### Description of the Lexical Resource

In this section, we describe the lexical resources that we used for our experiments on automatic extraction of lexical functions from raw text corpora.

**Compilation** Firstly, the Spanish Web Corpus [SpWC] was chosen as a source of verb-noun pairs with the pattern verb + direct object. All such verb-noun pairs used in the Spanish Web Corpus five or more times, were extracted automatically from the said corpus by the Sketch Engine [Kilgarriff *et al.* 2004], a web-based program for corpus processing. Fig. 1 displays the interface of the Sketch Engine where several corpora are listed including the Spanish Web Corpus. The obtained list contained 83,982 verb-noun pairs, and it was ranked by frequency.

Secondly, one thousand pairs were taken from the upper part of the list, i.e. most frequent verb-noun pairs.

Thirdly, in the list of one thousand pairs, erroneous combinations were marked with the label ERROR. Erroneous pairs included, for instance, past participle or infinitive instead of noun, or contained symbols like --, «, © instead of words. How did errors emerge? The automatic extraction procedure was set to search for combinations with the pattern verb + direct object in the corpus. This procedure needs part of speech (POS) and lemma information, and such data is supplied by TreeTagger, software used to annotate the Spanish Web Corpus with POS and lemmas. The TreeTagger is a leading tool applied for POS tagging and lemmatization, it achieves high accuracy but still is error-prone. Due to errors made by the TreeTagger, the set of extracted verb-noun pairs contained fallacious combinations. For the sake of preserving the original design of automatically extracted set, these incorrect combinations were not removed from the list but identified as wrong. The total number of erroneous pairs was 61, so after their removal the list contained 939 pairs.

Fourthly, collocational verb-noun pairs were annotated with lexical functions. The rest of the pairs were annotated as free word combinations using the label FWC.

Lastly, all verbs and nouns in the list were disambiguated with word senses from the Spanish WordNet, an electronic lexicon structured the same way as WordNet for English. For some verb-noun pairs, relevant senses were not found in the above-mentioned dictionary, and the number of such pairs was 39. For example, in the combination *dar cuenta, give account*, the noun *cuenta* means *razón, satisfacción de algo (reason, satisfaction of something)*. This sense of *cuenta* is taken from Diccionario de la Lengua Española (Dictionary of the Spanish Language) [RAE 2001]. Unfortunately, this sense is absent in the Spanish WordNet so the expression *dar cuenta* was left without sense annotation. All such words were annotation N/A, i.e. not available.

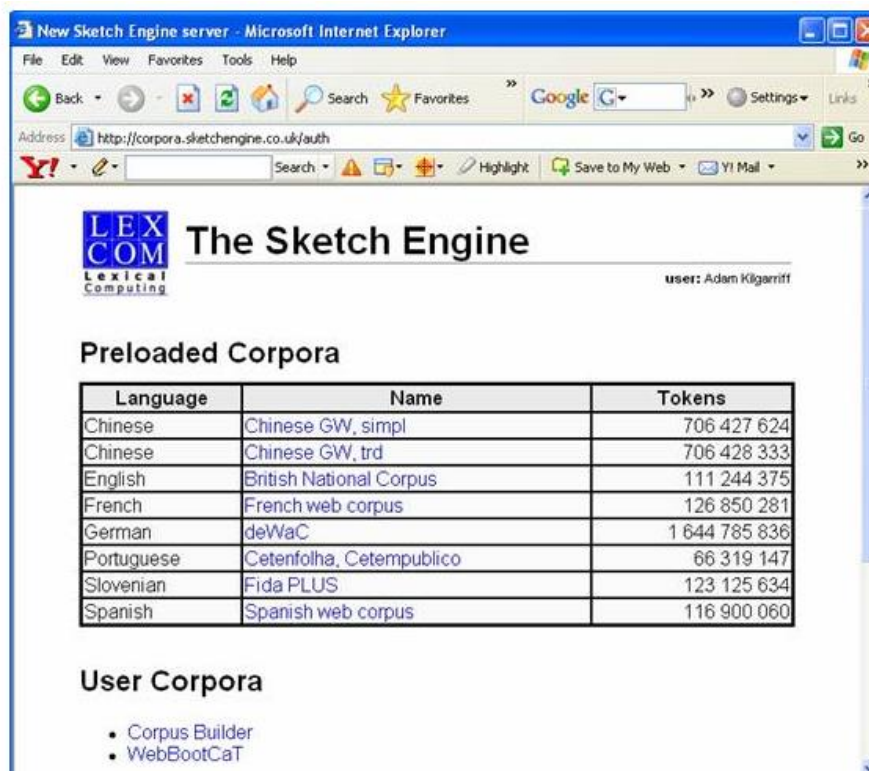


Fig 1. Sketch Engine with the Spanish Web Corpus.

The annotated list was formatted as a table and saved in an MS Excel file. Fig. 2 shows the process of the compilation of the lexical resource schematically.

**Contents of the lexical resource** A partial representation of the list is given in Table 5; Table 6 lists all lexical functions found in the list of 1000 most frequent verb-noun

pairs, their frequencies in the Spanish Web Corpus, and the number of examples for each of them.

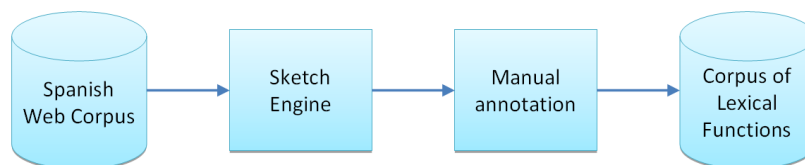


Fig. 2. The process of lexical resource compilation.

### 3.2 Data for the Test Sets

To build the test set, we extracted all verb-noun pairs from a corpus other than the corpus used to construct the training sets. So the data for test sets was mined from the Spanish Treebank Cast3LB [Civit and Martí 2004]. The number of all verb-noun pairs extracted from Cast3LB was 5181. We constructed four test sets, including, respectively, 100%, 75%, 50%, and 25% of all verb-noun pairs taken from Treebank Cast3LB.

Table 5. Partial representation of the lexical resource

LF/ FWC/ ERROR	VERB	Verb Sense Number	NOUN	Noun Sense Number	FREQ
Oper <sub>1</sub>	dar	2	cuenta	N/A	9236
CausFunc <sub>0</sub>	formar	2	parte	1	7454
Oper <sub>1</sub>	tener	1	lugar	4	6680
Oper <sub>1</sub>	tener	1	derecho	1	5255
CausFunc <sub>1</sub>	hacer	2	falta	N/A	4827
CausFunc <sub>1</sub>	dar	9	lugar	4	4180
Oper <sub>1</sub>	hacer	15	referencia	2	3252
Func <sub>0</sub>	hacer	N/A	año	2	3211
Oper <sub>1</sub>	tener	1	problema	7	3075
Func <sub>0</sub>	hacer	N/A	tiempo	1	3059
IncepOper <sub>1</sub>	tomar	4	decisión	2	2781
Oper <sub>1</sub>	tener	1	acceso	3	2773
Oper <sub>1</sub>	tener	1	razón	2	2768
Caus <sub>2</sub> Func <sub>1</sub>	llamar	8	atención	1	2698
Oper <sub>1</sub>	tener	1	sentido	1	2563
ERROR	haber	ERROR	estado	ERROR	2430
FWC	hacer	6	cosa	3	2374
Oper <sub>1</sub>	tener	3	miedo	1	2226
ERROR	haber	ERROR	hecho	ERROR	2168

We did not disambiguate verb-noun pairs for the test sets manually. Instead, for each verb-noun, we built all possible verb-noun combinations of all senses in the Spanish WordNet. As an example, let us consider the pair *representar papel*, lit. *represent role*. The verb *representar* has 12 senses in the Spanish WordNet, and the noun *papel*, 5. This gives totally 60 combinations of *representar* and *papel* (12 multiplied by 5). The initial list for the test set included 5,181 verb-noun pairs, which resulted in totally 96,079 instances in the test set. A partial representation of the list is given in Fig. 3.

**Table 6.** Lexical functions with their respective frequency in corpus and the number of instances in the list of verb-noun pairs

LF	Freq	#	LF	Freq	#
Oper <sub>1</sub>	165319	280	PerfFunc <sub>0</sub>	1293	1
FWC	70211	202	Caus <sub>1</sub> Oper <sub>1</sub>	1280	2
CausFunc <sub>1</sub>	45688	90	Caus <sub>1</sub> Func <sub>1</sub>	1085	3
CausFunc <sub>0</sub>	40717	112	IncepFunc <sub>0</sub>	1052	3
ERROR	26316	61	PermOper <sub>1</sub>	910	3
Real <sub>1</sub>	19191	61	CausManifFunc <sub>0</sub>	788	2
Func <sub>0</sub>	17393	25	CausMinusFunc <sub>0</sub>	746	3
IncepOper <sub>1</sub>	11805	25	Oper <sub>3</sub>	520	1
Oper <sub>2</sub>	8967	30	LiquFunc <sub>0</sub>	514	2
Caus <sub>2</sub> Func <sub>1</sub>	8242	16	IncepReal <sub>1</sub>	437	2
ContOper <sub>1</sub>	5354	16	Real <sub>3</sub>	381	1
Manif	3339	13	PlusOper <sub>1</sub>	370	1
Copul	2345	9	CausPerfFunc <sub>0</sub>	290	1
CausPlusFunc <sub>0</sub>	2203	7	AntiReal <sub>3</sub>	284	1
Func <sub>1</sub>	1848	4	MinusReal <sub>1</sub>	265	1
PerfOper <sub>1</sub>	1736	4	AntiPermOper <sub>1</sub>	258	1
CausPlusFunc <sub>1</sub>	1548	5	ManifFunc <sub>0</sub>	240	1
Real <sub>2</sub>	1547	3	CausMinusFunc <sub>1</sub>	229	1
FinOper <sub>1</sub>	1476	6	FinFunc <sub>0</sub>	178	1

```

v tener 1          n aire 11          n error 3
n aire 1          v tener 9          v salir 1
v tener 1          n aire 12          n error 4
n aire 2          v salir 1          v salir 1
...              n error 1          n error 5
v tener 9          v salir 1          v salir 1
n aire 10         n error 2          n error 6
v tener 9          v salir 1          ...
    
```

**Fig. 3.** A part of the list of verb-noun pairs used for building the test set.

## 4 A New Method of Meaning Representation

In this section, we present our novel method for representing the meaning of words suitable for mining for lexical functions.

#### 4.1 Data Representation

Each verb-noun pair in the training set and in the test set is represented as a set of all hypernyms of the noun and all hypernyms of the verb. The noun and the verb of the verb-noun pair were considered as zero-level hypernyms and thus were included in the set of hypernyms.

##### Hypernyms and Hyponyms

In linguistics, a hyponym is a word or phrase whose meaning is included within the meaning of another word, its hypernym. To put it simpler, a hyponym shares a type-of relationship with its hypernym. For example, *restaurant*, *rest house*, *planetarium*, *observatory*, *packinghouse*, *outbuilding*, *Pentagon* are all hyponyms of *building* (their hypernym), which is, in turn, a hyponym of *construction*.

In computer science, the relationship of hypernymy is often termed an "is-a" relationship. For example, the phrase *Restaurant is a building* can be used to describe the hyponymic relationship between *restaurant* and *building*.

Thus, hypernymy is the semantic relation in which one word is the hypernym of another one.

##### Spanish WordNet as a Source of Hypernyms

The Spanish WordNet follows the EuroWordNet [Vossen 1998] framework and is structured in the same way as the American WordNet for English [Miller 1998], namely, in terms of synsets (sets of synonymous words) with basic semantic relations between them.

Spanish nouns and verbs are organized into synonym sets, each representing one underlying lexical concept. Different relations, for example, hypernym relations, link the synonym sets.

Since all verbs and nouns have been disambiguated, hypernyms can be found for each word that has been annotated with its sense of the Spanish WordNet [SpWN]. Hypernyms were extracted automatically from the database of the dictionary referenced above. Fig. 4 and Fig. 5 display the interface of the Spanish WordNet as it is seen on the web. In the interface, we see hypernyms of *gato* "cat".

##### Hypernyms as a Meaning Representation

A difference between data representation in our experiments and data sets used in [Wanner *et al.* 2006] should be noted here. In the paper just referenced, every word in the training set was accompanied by its synonyms and hypernyms, its own Base Concepts (BC) and the BCs of its hypernyms, its own Top Concepts (TC) and the TCs of its hypernyms taken from the Spanish part of the EuroWordNet [Vossen 1998]. Unlike those experiments, in our work for the sake of simplicity we included only hypernyms in our training sets.

Though in this case the data is annotated with less information, i.e. only with hypernyms, or in other words, only hypernyms are used to represent the meaning of verb-noun pairs, we hope that hypernyms would be sufficient to distinguish between lexical functions. Up to now, there has not been any research done that compares different data

representations for the task of predicting lexical functions of verb-noun pairs. Here we can remember the original intent of WordNet compilers [Miller 1998] who claimed that the meaning of any word could be described sufficiently well for at least human understanding by semantic relations only, like “is-a-kind-of” semantic relation of hypernym hierarchy.

Later, the authors of WordNet admitted that their previous assumption had been wrong and glosses were added to distinguish synonym sets. Though practical significance of glosses is generally accepted, we intent to study how well the meaning of lexical functions can be distinguished if only hypernym information is taken into account.

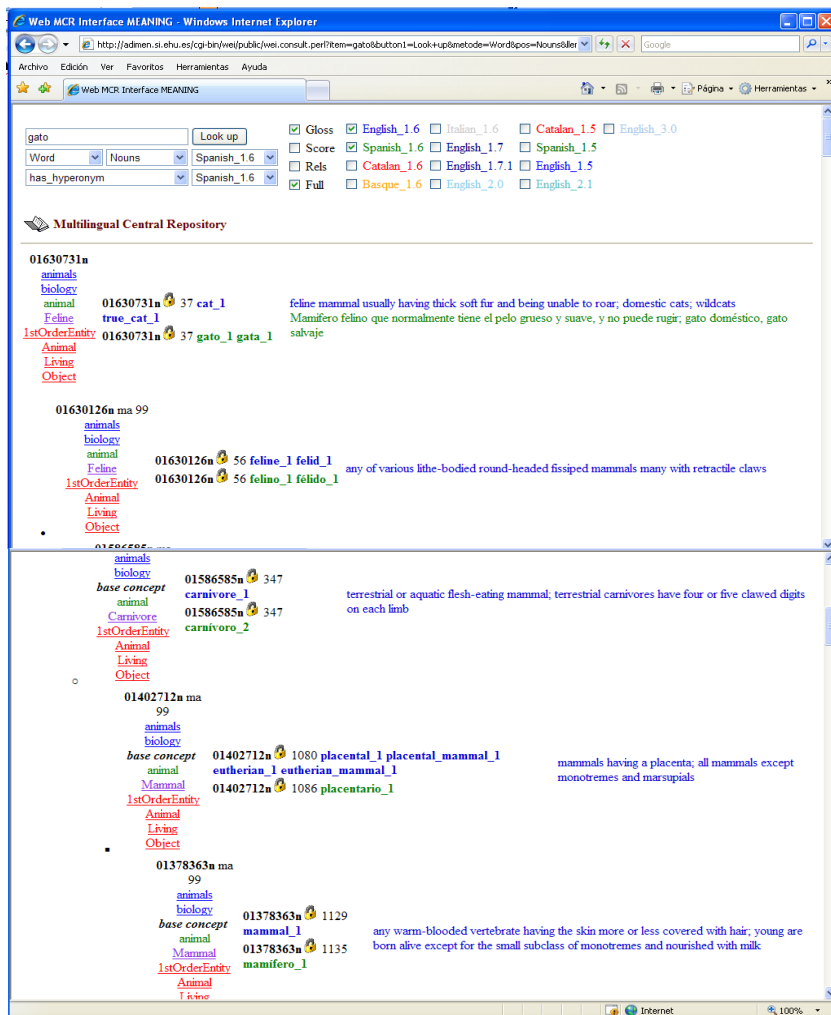


Fig. 4. The Spanish WordNet, hyperonyms for *gato*, cat.

Further research is needed to investigate how information other than hypernym taxonomy, for example, that of semantic ontologies, changes the performance of machine learning algorithms.

## 4.2 Linguistic Description of Training Sets and Test Sets

### Lexical Functions Chosen for Experiments

Our choice of lexical functions depends on the number of examples that each lexical function has in the lexical resource of Spanish lexical functions created by us and described in Section 4.1. We have selected LFs that have the number of examples sufficient for machine learning experiments. [Wanner 2004] and [Wanner *et al.*2006] experimented with the following number of LF examples: the biggest number of examples that this researcher had in the training set was 87 for Oper<sub>1</sub> and the least number of examples was 33 for Oper<sub>2</sub>.

**Table 7.** Lexical functions chosen for the experiments

LF and # of examples	Meaning	Collocation: LF value + keyword	
		Spanish	English translation
Oper <sub>1</sub> 280	Lat. <i>operare</i> – ‘to do, perform’. Experience (if K is an emotion), carry out K.	<i>alcanzar un objetivo</i> <i>aplicar una medida</i> <i>corregir un error</i> <i>satisfacer una necesidad</i>	<i>achieve a goal</i> <i>apply a measure</i> <i>correct a mistake</i> <i>satisfy a necessity</i>
CausFunc <sub>0</sub> 112	Lat. <i>causare</i> – ‘to cause’. Do something so that K begins occurring.	<i>encontrar respuesta</i> <i>establecer un sistema</i> <i>hacer campaña producir un efecto</i>	<i>find an answer</i> <i>establish a system</i> <i>conduct a campaign produce an effect</i>
CausFunc <sub>1</sub> 90	A person/object, different from the agent of K, does something so that K occurs and has effect on the agent of K.	<i>abrir camino</i> <i>causar daño</i> <i>dar respuesta</i> <i>producir un cambio</i>	<i>open the way</i> <i>cause damage</i> <i>give an answer</i> <i>produce a change</i>
Real <sub>1</sub> 61	Lat. <i>realis</i> – ‘real’. To fulfill the requirement of K, to act according to K.	<i>contestar una pregunta</i> <i>cumplir el requisito</i> <i>solucionar un problema</i> <i>utilizar la tecnología</i>	<i>answer a question</i> <i>fulfill the requirement</i> <i>solve a problem</i> <i>use technology</i>
Func <sub>0</sub> 25	Lat. <i>functionare</i> – ‘to function’. K exists, takes place, occurs.	<i>el tiempo pasa</i> <i>hace un mes</i> <i>una posibilidad cabe</i> <i>la razón existe</i>	<i>time flies</i> <i>a month ago</i> <i>there is a possibility</i> <i>the reason exists</i>
Oper <sub>2</sub> 30	Undergo K, be source of K	<i>aprender una lección</i> <i>obtener una respuesta</i> <i>recibir ayuda</i> <i>sufrir un cambio</i>	<i>learn a lesson</i> <i>get an answer</i> <i>receive help</i> <i>suffer a change</i>



IncepOper <sub>1</sub> 25	Lat. <i>incipere</i> – ‘to begin’. Begin to do, perform, experience, carry out K.	<i>adoptar una actitud</i> <i>cobrar importancia</i> <i>iniciar una sesión</i> <i>tomar posición</i>	<i>take an attitude</i> <i>acquire importance</i> <i>start a session</i> <i>obtain a position</i>
ContOper <sub>1</sub> 16	Lat. <i>continuare</i> – ‘to continue’. Continue to do, perform, experience, carry out K.	<i>guardar silencio</i> <i>mantener el equilibrio</i> <i>seguir un modelo</i> <i>llevar una vida (ocupada)</i>	<i>keep silence</i> <i>keep one’s balance</i> <i>follow an example</i> <i>lead a (busy) life</i>

Table 7 presents LFs that we have chosen for our experiments. For each LF, we give the number of examples, its meaning, and sample verb-noun combinations.

In the lexical resource, we have annotated free word combinations with the tag FWC. The number of FWC is 261. We considered free word combinations as a lexical function FWC in its own right and experimented how machine-learning algorithms can predict this class of word combinations. Therefore, the total number of LFs we experimented with is nine.

Remember, that in the training set and test set, each verb-noun combination is represented as a set of all hypernyms of the noun and all hypernyms of the verb. To construct this representation, the number of sense for every verb and noun must be identified. However, sometimes, an appropriate sense was absent in the Spanish WordNet. Such words were tagged with abbreviation N/A (not available) instead of the number of word sense. In the training set, we included only verb-noun combinations that are disambiguated with word senses of the Spanish WordNet. In Table 8, the numbers of examples include only the verb-noun pairs in which all the words are disambiguated with the Spanish WordNet.

**Table 8.** Number of verb-noun combination in the test sets

Test set	Number of verb-noun combinations
100%	5181
75%	3886
50%	2590
25%	1295

The total number of examples for all 9 lexical functions is 900.

### Training Sets

For each of 9 LF chosen for experiments, we built a training set, so we had 9 training sets. All training sets included the same list of 900 verb-noun combinations. The only difference between training sets was the annotation of examples as positive and negative. As an example, let us consider the training set for Oper<sub>1</sub>. In the list of 900 verb-noun pairs, there are 266 examples of Oper<sub>1</sub>, so these examples are marked as positive in the training set, and all the rest of verb-noun combinations whose number is 634 (900

– 266 = 634) were marked as negative examples. This procedure was applied to each training set.

### **Test Sets**

The test sets were built independently of the training set. For this, 5181 verb-noun combinations for the test set were extracted from the Spanish Treebank Cast3LB [Civit and Martí 2004]. Four test sets were constructed, including, respectively, 100%, 75%, 50%, and 25% of all verb-noun pairs taken from Treebank Cast3LB. Words in the test set were not annotated with lexical functions. Table 9 gives the number of verb-noun pairs in all four test sets.

## **5 Conclusions**

Lexical functions represent important linguistic information. Their understanding is important for correct interpretation of texts by a person or a computer. However, manual compilation of the corresponding dictionaries is a tedious and costly work. What is more, as any linguistic phenomenon, they depend on language, thematic domain, and genre, and they can change with time.

Automatic acquisition of lexical functions from unstructured raw texts greatly alleviates the problem of compilation of combinatorial dictionaries and in particular dictionaries of lexical functions. In this paper, we have presented a detailed discussion of the very notion of collocation and a review of existing approaches to automatic acquisition of lexical functions. Then we presented our methodology for representing the meaning of words via sets of direct and indirect hypernyms. Such a representation is particularly useful for automatic extraction of lexical functions from unprepared raw text corpora.

In our future work, we will apply this semantic representation to actual compilation of large dictionaries of lexical functions via supervised learning of semantic similarity measure between words.

## **References**

1. [Alonso Ramos 2003] Alonso Ramos, M.: Hacia un Diccionario de colocaciones del español y su codificación. In: M. A. Martí et al. (eds.), *Lexicografía computacional y semántica*. Barcelona: Edicions de l'Universitat de Barcelona, pp. 11–34.
2. [Alonso Ramos et al. 2008] Alonso Ramos, M., Rambow O., Wanner L.: Using semantically annotated corpora to build collocation resources. *Proceedings of LREC, Marrakesh, Morocco*, pp. 1154–1158.
3. [Apresjan 2004] Apresjan, Ju. D.: About semantic nonemptiness and motivatedness of verbal lexical functions. (In Russian.) *Voprosy jazykoznanija*: pp. 3–18
4. [Benson 1986] Benson, M., Benson, E. and Ilson R. 1986. *The BBI Combinatory Dictionary of English*. John Benjamins, Amsterdam.
5. [Benson 1990] Benson, M. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1), 23-35.

6. [Bolshakov and Gelbukh 1998] I. Bolshakov, A. Gelbukh. Lexical functions in Spanish. Proc. CIC-98, Simposium Internacional de Computación, November 11-13, 1998, Mexico D.F., pp. 383-395.
7. [Bolshakov and Gelbukh 2000] I.A. Bolshakov and A.F. Gelbukh. Classification of Collocations in a Lexical Database by Meaning of the Combined Words. In: Selected papers CIC-1999, CIC, IPN, Mexico City, 2000, pp. 5-15.
8. [Bolshakov and Gelbukh 2001] Igor A. Bolshakov and Alexander F. Gelbukh. A Large Database of Collocations and Semantic References: Interlingual Applications. *International Journal of Translation*, Vol.13, No.1-2, 2001, pp. 167-187.
9. [Bolshakov and Gelbukh 2002] Igor Bolshakov, Alexander Gelbukh. Word Combinations as an important part of modern electronic dictionaries. *Procesamiento de Lenguaje Natural*, No 29, 2002, Spain, p. 47-54.
10. [Bolshakov and Gelbukh 2004] I.A. Bolshakov, A. Gelbukh. Computational linguistics: models, resources, applications. IPN - UNAM - Fondo de Cultura Económica, Mexico, 187 pp.
11. [CEB] Concordancia electrónica de la Biblia Reina Valera 1960 online, <http://www.concordancia.bravefire.com/concordancia.php/>, last viewed on June 07, 2010
12. [Civit and Martí 2004] Civit, M., Martí, M.A.: Building Cast3LB: A Spanish Treebank. In: *Research on Language and Computation*, vol. 2(4), pp. 549–574. Springer, Netherlands
13. [Cowie 1994] Cowie, A. P. 1994. Phraseology. In Asher, R. E. (ed.). *The Encyclopedia of Language and Linguistics*. Oxford, Pergamon Press.
14. [DiCE] Diccionario de colocaciones del Español, <http://www.dicesp.com/paginas/>, last viewed June 08, 2010
15. [Firth 1957] Firth, J. R.: Modes of Meaning. In J. R. Firth, *Papers in Linguistics 1934–1951* (pp. 190–215). Oxford: Oxford University Press.
16. [Fontecha 1941] Fontecha, C.: *Glosario de voces comentadas en ediciones de textos clásicos*. Madrid: CSIC
17. [Fontenelle 1994] Fontenelle, T. 1994. What on Earth are Collocations? *English Today* 10:4, 42-48.
18. [Gelbukh and Sidorov] A. Gelbukh, G. Sidorov. *Procesamiento automático del español con enfoque en recursos léxicos grandes*. IPN, Mexico, 2006, 240 pp.
19. [Gledhill 2000] Gledhill, C. J. 2000. *Collocations in Science Writing*. Tübingen, Gunten Narr Verlag.
20. [Halliday 1961] Halliday, M. A. K.: Categories of the Theory of Grammar. *Word* 17, 241–292.
21. [Hausmann 1984] Hausmann, F. J. 1984. Wortschatzlernen ist Kollokationslernen. *Zum Lehren und Lernen französischer Wortverbindungen. Praxis des neusprachlichen Unterrichts*, 31, pp. 395 – 406.
22. [Howarth 1996] Howarth, P. 1996. *Phraseology in English academic writing. Some implications for language learning and dictionary making*. Tübingen, Niemeyer.
23. [Kilgarriff et al. 2004] Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D.: The Sketch Engine. In *Proceedings of EURALEX*. France, Université de Bretagne Sud: pp. 105–116
24. [McVey and Wegmann 2001] McVey Gill, M., Wegmann, B.: *Streetwise Spanish Dictionary/Thesaurus*. Chicago: McGraw-Hill.
25. [Mel’čuk 1996] Mel’čuk, I. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In: Wanner, L. (Ed.) *Lexical Functions in Lexicography and Natural Language Processing*, pp.37-102. John Benjamin Publishing Company.

26. [Mel'čuk and Zholkovskij 1984] Mel'čuk, I. A. and Zholkovskij, A. K. 1984. An Explanatory Combinatorial Dictionary of the Contemporary Russian Language. Wiener Slawistischer Almanach, Sonderband 14, 1984.
27. [Mel'čuk et al. 1984] Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Adèle Lessard. Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques I. Les Presses de l'Université de Montréal, 1984.
28. [Mel'čuk et al. 1988] Mel'čuk, Igor, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, Suzanne Mantha. 1988. Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II. Les Presses de l'Université de Montréal.
29. [Miller 1998] Miller, G.A: Foreword. In: Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, pp. xv–xxii. MIT Press, Cambridge, Mass. (1998)
30. [Nastase and Szpakowicz 2003] V. Nastase and S. Szpakowicz. Exploring noun-modifier semantic relations. In Fifth International Workshop on Computational Semantics (IWCS-5), Tilburg, The Netherlands, pages 285–301, 2003.
31. [Nastase et al. 2006] Nastase, V., J. Sayyad-Shiarabad, M. Sokolova, and S. Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference. AAAI Press.
32. [Oelschläger 1940] Oelschläger, V.R.B. 1940. A Medieval Spanish Word-List: A Preliminary Dated Vocabulary of First Appearances Up To Berceo. Madison, Wisc.: University of Wisconsin Press.
33. [RAE 2001] Real Academia Española 2001. Diccionario de la Lengua Española. (Twenty Second Edition.) Madrid: Real Academia Española
34. [Ruppenhofer et al. 2006] Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. R. and Scheffczyk, J. 2006. FrameNet II: Extended Theory and Practice. Available at <http://framenet.icsi.berkeley.edu/book/book.pdf>. ICSI Berkeley
35. [Sanromán 1998] B. Sanromán. Contribución lexicográfica al estudio de los nombres de emoción. Master's thesis, Universidad de Coruña.
36. [Sanromán 2003] B. Sanromán. Semántica, sintaxis y combinatoria léxica de los nombres de emoción en español. PhD thesis, Helsinki: University of Helsinki.
37. [Sinclair et al. 2004] J. Sinclair, S. Jones, R. Daley. English collocation studies: The OSTI report. Continuum. 2004.
38. [SpWC] Spanish Web Corpus in the Sketch Engine. 3 May 2010. <http://trac.sketchengine.co.uk/wiki/Corpora/SpanishWebCorpus>
39. [SpWN] Spanish WordNet, [http://www.lsi.upc.edu/~nlp/web/index.php?Itemid=57&id=31&option=com\\_content&task=view](http://www.lsi.upc.edu/~nlp/web/index.php?Itemid=57&id=31&option=com_content&task=view), last viewed June 02, 2010
40. [Van Roey 1990] French-English Contrastive Lexicology: An Introduction. Louvain-la-Neuve, Peeters.
41. [Vossen 1998] P. Vossen. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic, Dordrecht.
42. [Wanner 2004] Leo Wanner. Towards automatic fine-grained semantic classification of verb-noun collocations. Natural Language Engineering (2004), 10:2:95–143 Cambridge University Press.
43. [Wanner et al. 2006] Wanner, L., Bohnet, B. and Giereth, M. 2006. What is beyond Collocations? Insights from Machine Learning Experiments. EURALEX.