# Towards a Framework of Verbal Compositionality

Gaspar Ramírez[1], Héctor Jiménez-Salazar[2] & Jim Fidelholtz[1]

[1]Maestría en Ciencias del Lenguaje,
Casa Amarilla, 2 Oriente, CP 72000
[2]Facultad de Ciencias de la Computación,
Ciudad Universitaria, Edif. 135, CP 72570
Benemérita Universidad Autónoma de Puebla
Puebla, Pue., México.
{gasparfirst, hgimenezs}@gmail.com, jfidel@siu.buap.mx

**Abstract.** Lexical Semantic Theory faces problems of how words acquire different meanings in distinct contexts. In this work we analyze the first steps that should be taken in order to constitute a combinatory dictionary of Spanish verbs. Our proposal is conceived within the Generative Lexicon approach of James Pustejovsky, and we discuss some ideas of how to build a dictionary with such characteristics.

## 1 Introduction

In modern linguistics and Natural Language Processing (NLP), regular polysemy has recently become a phenomenon of natural languages which people from a wide variety of fields have begun to study. Pustejovsky [5] noticed that the specific polysemy of some aspectual verbs like *terminar, comenzar*, experiencer verbs like *disfrutar*, and many causatives (we have given their Spanish equivalents) depends largely on the particular context in which they occurred. Some examples are

(0)a. Juan terminó/ disfrutó su cigarro.
(0)b. Juan terminó/ disfrutó su café.

In sentence (0a) terminar/disfrutar means "*terminar/disfrutar de fumar*", whereas in (1b) it means "*terminar/disfrutar de beber*". Here *su cigarro* and *su café* 'coerce' the meaning of the verbs. Pustejovsky proposed a Generative Lexicon (GL) where categories such as verbs and nouns are linked by means of a flexible mechanism made up of different levels of semantic representation. This mechanism is able to capture the contribution of verb arguments to the meaning of the verb.

We think that what categories and elements of these categories refer to could be better understood from a co-compositional or co-occurrence perspective where new features of meaning may arise. Thus lexical and syntagmatic features form the core of our proposal for elaboration of a lexicon of verbs in Spanish.

This research is inserted within the field of lexical semantics. The theory of lexical semantics deals with the problem of how words can acquire different

© *A. Gelbukh, S. Torres, I. López (Eds.)*
*Advances in Artificial Intelligence*
*Research in Computing Science 20, 2006, pp. 39-48*

meanings in different contexts, how meanings can arise compositionally, and how semantic types can be mapped into the syntactic forms in a predictable way. To discover what factors in a speech act are responsible for our ability to convey the wealth and diversity of meaning with a quite limited number of linguistic resources is a task that is worthwhile, since empirical research in this area is scarce. The aim of lexical semantics is therefore to provide a detailed description of how expressions in language acquire their content and how this content seems to suffer continued modification and modulation in novel contexts. This research attempts to analyze lexical semantics of verbs both individually and in combination with other lexical items in order to incorporate their linguistically fine-grained description in a combinatory explicative dictionary (cf. Mel'čuk [4]) covering the following important considerations in formal semantic theory [8].

The methodology that will be employed consist, on the one hand, of grouping the meanings of verbs according to the syntactic frame in which they participate; this is commonly known as verb alternations. On the other hand, it also includes aspect or Aktionsarten [7] of verbs as a way to capture the way verbs are conceptualized. Cognitively, this feature is as important as the difference that exists among countable and uncountable nouns. Finally, verbs will be classified in semantically unique classes. Of course, there are some verbs which may appear in more than one class due to the kind of action which they perform.

In this work two theoretical assumptions are considered to describe in detail the semantics and lexicon of any natural language. First, it is known that without considering the syntactic structure of a language, the study of lexical semantics will not work. In other words, there is no way to separate completely the meaning from the structure that carries it. The second assumption also says that the meanings of words should somehow reflect the deeper conceptual structures in the cognitive system, as well as the domain it operates in.

From a computational lexical semantic perspective the following principles should in some way be considered. First, a clear notion of well-formedness in semantics will be necessary in order to characterize a theory of word meaning. Secondly, lexical semantics should look for representations richer than the descriptions from thematic roles. Thirdly, several levels of semantic interpretation should be used [9].

Recent works in lexical semantics have been largely focused on clarifying the nature of verb classes and the syntactic structure that each allows (cf. Levin 1985, 1993, taken from [5]). However, we should explain syntagmatically why verb classes behave as they do, and what consequences these distinctions have for the rest of the lexicon and grammar. Thus the aim of this research is to identify the similarities and differences, semantic as well as non-semantic, of verbs considered compositionally, according to the context in which they occur Following Pustejovsky [5], a lexical semantic theory should not merely map the number of lexical meanings per sentence, on an individual basis. Rather, it should capture the relations between words in a way which facilitates this mapping.

In order to support the lexical representation proposed by Pustejovsky, two basic concepts and their use are introduced in Section 2 and 3 of this paper. Section 4 presents a discussion on the development of our combinatory dictionary, referring to some related works.

## 2 Semantic Classes and Categorial Alternations

In the tradition of formal semantics, perhaps the most relevant aspect of the meaning of a word is its semantic type. Therefore type or categorial information determines not only how a word behaves syntactically, but also what the elements of such categories refer to.

Some examples follow. The verbs *amar* and *odiar* may be considered as relations among individuals in the world, whereas *mujer* would select the set of all individuals that are women. As type distinctions are generally very broad, lexical semantics distinguishes even selectional subsets for members of these categories. A finer lexical semantic representation for the lexical items and its combination with other item is then necessary in order to characterize broadly the expressive power of languages.

### 2.1 Semantic Classes

This research is based on one of the oldest semantic classifications of verbs, the aspectual class or *Aktionsarten.* This classification considers that verbs and verbal phrases vary according to the types of events that they denote in the world or, in other words, the kind of action they denote. It is usually assumed that there are at least three aspectual types: state, activity, and event, where the last sometimes is divided into accomplishment and achievement events.

Some examples show what we mean by aspectual class. The verb *caminar* in sentence (1) denotes an activity of unbounded duration, that is, the sentence itself does not carry information about the temporal extension of the activity, although deictically it turns out to be an event that did finished in the past.

(1) María caminó ayer.
(2) Maria caminó a su casa ayer.

It is said that sentence (1) denotes an activity. Other examples of this class of verbs are: *dormir, correr, trabajar, beber*, etc. On the other hand, sentence (2) also conveys the same information as the previous one, except that in this case the constraint appears that María finished walking when she arrives to her house. Although there isnt any explicit reference to duration of the activity, this sentence states that the process has a logical culmination, since the activity finishes when María gets home. It is said that this kind of sentence denotes an accomplishment event.

Just as the verb *caminar* seems by default to represent an activity in lexical terms, there are verbs that seem to denote accomplishments lexically. For example, the verbs *construir* and *destruir*, in their typical transitive use, denote accomplishment events since there is a logical culmination to the activity performed.

(3) María construyó una casa.
(4) María destruyó la mesa.

In sentence (3) the coming into being of the house is the culmination of María's act, while in (4) the non-existence of something referred to as a table is the direct culmination or consequence of this act. Verbs of creation are the best examples of accomplishments events. One of the classical diagnostics to probe if a verb (phrasal or not) denotes an accomplishment is its modification by time adverbials like *en una hora*, that is, so-called adverbial frames. Notice that both derived accomplishments (5) and lexical accomplishments (6) permit this modification, while activities (7 and 8) do not.

(5) María caminó a la tienda en una hora.
(6) María construyó la casa en un año.
(7) *Juan se bebió en 20 minutos.
(8) *María se trabajó en una hora.

Apparently, an adverbial frame requires that the verb or phrasal verb make an explicit reference to a change of state, a precondition which is missing in (7) and (8).

An achievement, on the other hand, is an event that undergoes a change of state, similarly to what happens in an accomplishment event, but where the change is thought of as occurring instantaneously. For example, in sentences (9), (10) and (11) the change is not gradual, but something that has a point-like character. Therefore, modification with punctual adverbials such as *a las 3 en punto* suggests that the sentence denotes an achievement event.

(9) Juan murió a las 3 en punto.
(10) Juan encontró su cartera a las 3 en punto.
(11) Maria llegó a la media noche.

Of course, punctual adverbial modification is not restricted just to achievement events, as the following examples show:

(12) Ella nadó el canal a las 10:00 a.m.
(13) El pianista ejecutó la sonata al medio día.
(14) Jaime enseñó su seminario de tres horas a las 2:30.
(15) Él dictó su conferencia a las 4 p.m.

Here the punctual adverbial indicates the beginning of an event with certain duration. It seems that some lexical proprieties of verbs may be affected by the sort of complement with which they interact.

As we can see by the examples given so far, the kind of event that a verb denotes may vary from a compositional perspective. Therefore co-occurrence meaning as well as compositionality should be considered when describing a lexical item. A shift of meaning in the verb arises as a result of the syntagmatic interactions and the semantic and syntactic relationship of the verb with the rest of the items in the sentence.

## 2.2 Verb Alternations

We also employ a recently developed methodology to group the meanings of verbs in semantic classes through the analysis of the syntactic frames in which they participate; that is, common grammatical verb alternations. Here we can mention as an example of similar work the MIT Lexicon Project, which outlines a large classification of argument verb alternations in English in order to classify verbs into semantically unique classes. Let us consider the following examples, the verbs *hundir, rodar*, and *romper* all have transitive and intransitive forms when their lexical senses are related to the interpretative characteristic of causation.

(16) a. El bote se hundió en un clima tormentoso.
b. El avión hundió el bote en un clima tormentoso.

(17) a. La pelota rodó por la colina.
b. Bill rodó la pelota por la colina.

(18) a. Súbitamente, la botella se rompió.
b. Súbitamente, Maria rompió la botella.

(19) a. La carta llegó a tiempo.
b. *El cartero llegó la carta a tiempo.

(20) a. Mi terminal murió anoche.
b. *La tormenta murió mi terminal anoche.

(21) a. La torre de bloc cayó.
b. *Zacarías cayó la torre de bloc.

Although sentences (19b), (20b), and (21b) are ill-formed, they are certainly understandable. A lexical semantic theory should specify what these two classes share; for example, both have intransitive grammatical forms. Thus it is important to identify similarities among verbs for establishing a domain where lexical items are somehow unified (unification), but equally important is the characterization of how verbs differ (individualization); for example, the latter group does not allow transitive form. The question is whether it is possible to identify the linguistically relevant features that lead us to the distinct behavior of the transitive verbs above. However, as Pustejosvky [5] claimed, we can only explain the behavior of a verb's semantic class can be achieved only by acknowledging that the syntactic patterns in an alternation are dependent on the information carried by the arguments in the patterns themselves. In other words, the diversity of complement types that a verb or other category may take is in large part determined by the semantics of the complements themselves.

There are other alternations of argument change than the ones discussed above, as well as alternations of argument drop.

(22)a. La mujer comió su cena rápidamente.
b. La mujer comió rápidamente.

(23)a. El perro devoró la galleta.
b. *El perro devoró.

(24)a. Juan bebió la cerveza febrilmente.
b. Juan bebió febrilmente.

(25)a. Febrilmente, Juan se hecho de un trago la cerveza.
b. *Juan se hecho de un trago febrilmente.

(26)a. María tarareó una canción mientras caminaba.
b. María tarareó mientras caminaba.

(27)a. María interpretó una canción mientras comía su cena.
b. *María interpretó mientras comía su cena.

Grammatical alternations, along with aspect or *Aktionsarten*, can be used throughout the grammar of a language to make semantic distinctions between verbs on the basis of syntactic behavior, and in the same sense to find similarities. Using categorial selection information as well as the data from grammatical alternations, verbs can be grouped in semantic classes which, at the same time, have predictable syntactic behavior.

## 3 Levels of Representation

Next, we explain how lexical information is organized within a GL.

Following Pustejovsky [5], a GL is regarded as a computational system that involves at least 4 levels of representation.

1. ARGUMENT STRUCTURE: Specification of number and type of logical arguments, and how they are realized syntactically.
2. EVENT STRUCTURE: definition of the event type of a lexical item and a phrase. Types include STATE, PROCESS, and TRANSITION, and events may have a subevent structure.
3. QUALIA STRUCTURE: Modes of explanation composed of FORMAL, CONSTITUTIVE, TELIC and AGENTIVE roles.
4. LEXICAL INHERITANCE STRUCTURE: Identification of how a lexical structure is related to other structures in the type lattice, and its contribution to the global organization of the lexicon.

Thus he argues that a set of generative devices connects these four levels, providing for the compositional interpretation of words in contexts. These devices are simply semantic transformations, all involving well-formedness conditions on type combinations.

- TYPE COERCION. Where a lexical item or phrase is coerced to a semantic interpretation by a governing item in the phrase, without changing of its syntactic type.
- SELECTIVE BINDING. Where a lexical item or phrase operates specifically on the structure of a phrase, without changing the overall type in the composition.
- CO-COMPOSITION. Where multiples elements within a phrase behave as functors, generating new non-lexicalized senses for the words in composition. This also includes cases of underspecified semantic forms becoming contextually enriched, such as manner co-composition, feature transcription, and light verb specification.

When we define the functional behaviour of lexical items at different levels of semantic representation, we hope to get at a characterization of the lexicon as an active and integral component in analyzing the compositional aspects of sentence meaning.

As we can see by the examples that we have presented so far, new meanings of words seem to emerge if words are regarded in composition rather than considering them as isolated and unrelated lexical items. Therefore, a generative lexicon must be seen as a structured system where different grammatical categories are linked in order to show the semantic relatedness which can arise within a co-occurrence and co-compositional semantic frame.

Next, we shall offer an example of a standard entries under our proposal based on a merger of a Combinatory Explicative Dictionary (CED) with a generative lexicon. In addition to the semantic type system, we also include other items needed in a detailed description of an entry. These items are: the meaning zone, the co-occurrence constraints zone, and the zone of illustrations. Here we offer the meaning of the word in Spanish, with its English translation. However, in this work the zone of co-occurrence constraints is not yet taken up.

### Standard definitions

> **Construir** v. tr. (lat. Construere) [29]. Hacer una obra material o inmaterial, juntando los elementos de acuerdo a un plan: *construir un edificio, construir una teoría*, 2. LING. Ordenar y enlazar debidamente las palabras en la oración o frase 3. MAT. Trazar o construir un polígono.
> [**Construir** v. tr. (lat. Construere) [29]. Make a concrete or abstract work joining the elements according to a plan: *construct a building, construct a theory.* 2. LINGUISTICS. Order and correctly connect the words in a sentence or phrase. 3. MATHEMATICS. Plot or construct a polygon.]
> **Construir**: Crear una cosa material o inmaterial, agrupando las partes según un plan trazado.
> [**Construir**: Create a concrete or abstract thing, grouping the parts according to a plan.]

**Combined CED and GL**

$$\left[\begin{array}{lll}
\text{Construir} & & [\textit{Construct}] \\
\text{EVENTRSTR} & = & \left[\begin{array}{ll} \text{E1=} & \textbf{process} \\ \text{E2=} & \textbf{state} \\ \text{REST=} & \leftarrow \\ \text{HEAD=} & e_1 \end{array}\right] \\
\text{ARGSTR} & = & \left[\begin{array}{ll}
\text{ARG1=} & 1\left[\begin{array}{ll}\textbf{animate-individual} & \\ \text{FORMAL=} & \textbf{physobj}\end{array}\right] \\
X = 1; & \text{who constructs?} \\
\%\ \textit{Juan} \sim & [\%\ \textit{John} \sim] \\
\text{ARG2=} & 2\left[\begin{array}{ll}\textbf{entity} & \\ \text{CONST=} & 3 \\ \text{FORMAL=} & \textbf{physobj/absobj}\end{array}\right] \\
Y = 2; & \text{what?} \\
\% \sim \textit{una silla} \simeq & [\% \sim \textit{a chair} \simeq] \\
\% \sim \textit{una teoría} \simeq & [\% \sim \textit{a theory} \simeq \\
\text{D-ARG1} & 3\left[\begin{array}{ll}\textbf{material} & \\ \text{FORMAL=} & \textbf{mass}\end{array}\right] \\
Z = 3; & \text{from what?} \\
\% \simeq \textit{sobre el clima} & [\% \simeq \textit{about the climate}] \\
\% \simeq \textit{de madera} & [\% \simeq \textit{of wood}]
\end{array}\right] \\
\text{QUALIA} & = & \left[\begin{array}{ll} \textbf{create-lcp} & \\ \text{FORMAL=} & \textbf{exist}(e_2, 2) \\ \text{AGENTIVE=} & \textbf{build-act}(e_1, 1, 3) \end{array}\right]
\end{array}\right]$$

## 4 Discussion

We have established the most important items of GL theory. From a methodological viewpoint it is necessary to build several elements using NLP tools in order to construct a viable lexicon. We think the following steps are indicated:

1. Create a list of lemmas: $L = \{x_1, \ldots, x_N\}$.
2. For each $x_i \in L$:
   (a) Create an initial matrix for $x_i$, $m_i$; e.g. using the most frequent sense.
   (b) Obtain all synsets of $x_i$ from EuroWordnet: $S_i = \{s_1, \ldots, s_n\}$.
   (c) Extract a large quantity of sentences from a corpus; e.g. the web: $O_i = \{o_1, \ldots, o_m\}$.
   (d) Assign a sense from $S_i$ to each $o_j \in O_i$ and cluster the $o_j$ according to its sense. Let $C_i = \{c_1, \ldots, c_k\}$ be the clusters obtained.
   (e) Analyze each cluster $c_l$ and update the matrix $m_i$.

The most difficult task in the above procedure is word sense disambiguation, but the manual work that the above steps imply cannot be ignored. Certainly, there are tools that assign the sense to a word in a context, but they are not very precise. Likewise, the process has to be semiautomatic. It is important to take advantage of other approaches to building lexicons under GL theory.

There have been many projects to create lexicons following the GL proposal. In our work, we try to resolve several problems that have come up in previous works. In the following paragraphs, we give a brief outline of some work related to the construction of lexicons.

Before analyzing the proposals, some important issues must be highlighted. First, a speaker can efficiently create a "new" sense of a word in a given context. Second, if we proceed to build a lexicon based on the GL approach using a corpus, the rules will be able to do better with "new" word uses, insofar as the corpus is larger. Thus a limit to the creative understanding of new uses of words will be the size of the corpus. That is, the challenge is for rules in the GL to provide enough information to proceed when faced with new uses of words, which will obviously be easier with a larger corpus. Kilgarriff [3] refers to the last point.

In [3] Kilgarriff focuses on the power of the GL approach. His evaluation is centered in non-standard word uses, trying to answer whether such uses could be analysed by GL strategies. A non-standard use was defined as not fitting under any dictionary definition of the word in a particular dictionary. He found that from 41 instances of non-standard uses just 5% (two instances) were plausible candidates for the GL treatment. So, without intending to undermine GL analysis, he shows that the GL is suitable for only some lexical phenomena, not all.

Because building a lexical resource is time-consuming and costly, Ruimy et al. [6] report the development and experimental implementation of a combined methodology of knowledge transfer from a monolongual Italian lexical resource to a French semantic lexicon. This work follows the theoretical principles of GL theory. The main idea is from an Italian lexicon to semi-automatically inferr a similar annotated French lexicon. They used translation word pairs provided by bilingual dictionaries in order to assign semantic properties given by the Italian lexicon to word senses of French. Their approach takes as much advantage as possible of similarity between French and Italian; the cognate approach, based on regularities of some suffixes in both Italian and French. On the other hand, in the cases that the cognate-based method was not applicable, they used sense indicators taken from bilingual dictionaries. The success rate for such suffixed words was 95%. Still, the methodology did not prove very efficient in completing the lexicon. However, the authors are hopeful that the methodology used can be applied in similar cases.

Qualia structure is generally understood as a representational tool for expressing the componential aspect of word meaning. While FORMAL, CONSTITUTIVE, TELIC and AGENTIVE qualia roles provide the basic structuring of a semantic type, Busa et al. [2] introduce the notion of Extended Qualia Structure (EQS) in the framework of the development of large-scale lexical resources, the SIMPLE model. EQS is motivated by the fact that lexical items may share the same structural properties. EQS is achieved by decomposing a qualia role into subtypes of the role consistent with its interpretation. There are strong types which create a new type of qualia, and weak types which add information to a type without changing its nature. The authors created a library of templates that provide the constraints and conditions for a lexical to belong to a type. SIMPLE may be viewed as a template-based framework for lexicon development because each type is associated with a template which provides the well-formedness condition for a lexical item to be of a given type.

We have seen that the GL theory continues to develop and has an impact on lexicon building. Furthermore, several strategies for constructing the dictionary of Spanish verbs may be exploited, as the works on this topic suggest.

Of course, all of the above suggestions need to be tried out for Spanish verbs, to see if the theory-based suggestions pan out in practice. In any case, we belive that many of them will prove useful, as has been the case in earlier work on other languages. The master thesis of the first author will include many of the topics mentioned in this paper, and the results will be reported in due course.

## References

1. Bouillon, Pierrete & Federica Busa: *The Language of Word Meaning*, Cambridge University Press, 2001.
2. Busa, Federica; Nicoletta Calzolari & Alessandro Lenci: Generative Lexicon and the SIMPLE Model: Developping Semantic Resources for NLP, in [1], pp 333–349, 2001.
3. Adam Kilgarriff: Generative Lexicon Meets Corpus Data: The Case of Nonstandard Word Uses, in [1], pp 312–330, 2001.
4. Mel'čuk, Igor A.: *Dependency Syntax: Theory and Practice*, Albany, State University of New York Press, 1988.
5. Pustejovsky, James: *The Generative Lexicon*, MIT Press, 1995.
6. Ruimy, Nilda; Pierrette Bouillon & Bruno Cartoni: Inferring a Semantically Annotated Generative French Lexicon from an Italian Lexical Resource, in *Third International Workshop on Generative Approaches to the Lexicon*, pp 218–226, 2005.
7. Vendler, Zeno: *Linguistics in Philosophy*, Cornell University Press, Ithaca, 1976.
8. Gelbukh, Alexander and Grigori Sidorov. Automatic selection of defining vocabulary in an explanatory dictionary. *Lecture Notes in Computer Science* **2276**:300–303, Springer, 2002.
9. Gelbukh, Alexander, and Grigori Sidorov. Hacia la verificación de diccionarios explicativos asistidos por computadora. *Estudios de Lingüística Aplicada* **38**:89–108, 2003.