Mariona Taulé Delor
*Universitat de Barcelona*

M. Antònia Martí Antonín
*Universitat de Barcelona*

Irene Castellón Masalles
*Universitat de Barcelona*

# Towards a 'VRQS' Representation

## Abstract

In this paper we would like to present a proposal for the representation of verbal lexical meaning in a Lexical Knowledge Base. We start from the hypothesis that verbal lexical meaning is the sum of different levels of semantic description: Argument Structure, Event Structure, Selectional Restrictions and Compositional Semantics. We are going to base our talk on the representation of the Compositional Semantic level, the decomposition of the meaning in semantic components and how they are saturated in Spanish. We present three different types of saturation (argumental, morphological and 'understood') and we will try to demonstrate how these sorts of saturation have different syntactic consequences.

## 1. Introduction

The work reported is part of an extended more general lexical study developed in the framework of the Acquilex Project[1] one of whose basic goals is the construction of a Computational Lexicon in which the lexical entries are represented in a multilingual Lexical Knowledge Base (LKB) (Copestake 1992)[2] which uses a representation language based on unification. LKB objects are represented as typed feature structures that are hierarchically ordered and allow inheritance of information.The lexical entries are treated as lexical signs following the proposal of "Head–Driven Phrase Structure Grammar" (HPSG) (Pollard & Sag 1987–1992).

The main purpose of this work is the representation of a specific type of semantic information, viz. that referring to the semantic components or entities which constitute the verbal lexical meaning, and the way in which this sort of information (labeled <vrqs>) is encoded in the LKB. At this level of representation we propose a semantic classification of verbs according to the different semantic components of which they are composed. In this sense, we start from the hypothesis that verbal lexical meaning is the sum total of different levels of semantic description that contribute to the whole verbal lexical meaning each in their own way (Pustejovsky 1991): Argument Structure, Event Structure, Selectional Restrictions and Compositional

Semantics.[3] Each one of these semantic levels contributes different types of information to the verbal meaning and, therefore, in order to give a complete semantic characterization of verbs the representation of all of them is necessary.

In this work we deal with the semantic subset of motion verbs in Spanish in order to illustrate the treatment and encoding of the Compositional Semantic level.

## 2. Verbs in the Lexical Knowledge Base

Verbal lexical entries are represented in the LKB as lexical signs in which morpho–syntactic and semantic information is encoded together (See Figure 1). Verbs are treated as 'head' elements of their sentences. In this sense, verbs provide the relational and semantic structure for the sentences in which they play the central role.

| lex–sign (sign) | verb–sign (lex–sign complex–sign) |
|---|---|
| <orth> = orth | <orth> = orth |
| <cat> = cat | <cat> = complex–cat |
| <sem> = sem | <sem> = verb–sem |
| <rqs> = rqs | <rqs> = vrqs. |
| <sense–id> = sense–id[4] | |

Figure 1: Lexical and verbal signs.

Argument Structure, Event Structure and the specification of Selectional Restrictions are encoded in the <verb–sem> feature, while the information concerning the different semantic components is defined in the <vrqs> feature. The semantic information included in <verb–sem> is represented as a conjunctive logical form in which the verbal predicate is characterized as a type of event and the semantic relation between verbal predicate and its arguments is expressed by means of thematic proto–roles. In the logical form, the selectional restrictions that can flesh out the arguments of the verbal predicate are specified as well.[5] The last level of semantic representation, encoded in the LKB as <vrqs>, includes information concerning the semantic class to which the verb belongs and also information about the verbal arguments saturated by the verb.

In the verbal sign, syntactic and semantic information are related by means of the coindexing of the subcategorized arguments specified in <cat> and the arguments of the logical form specified in <verb–sem>, in such a way that the verbal argument selection is evident.

Next we present, in a more detailed way, the nature of compositional representation, and we argue why such a level of representation is needed.

### 3. 'VRQS' and decomposition

The level of compositional representation includes the abstract specification of the meaning of a concrete verb by means of its decomposition into semantic components or entities (Talmy 1985; Jackendoff 1990). Each particular semantic domain is characterized by a defined subset of semantic components. Basing ourselves on the subset of motion verbs dealt with in our work, we illustrate the way we carried out the decomposition of meaning.

The abstract representation of the meaning of a motion verb (e.g.: *mover* (to move), *andar* (to walk), *subir* (to go up), *bajar* (to go down), *zigzaguear* (to zigzag), *gatear* (to crawl), *aletear* (to wing), etc.) involves, basically, Agent, Patient, Path, Manner and Motion as semantic components. In this area, it is necessary to distinguish verbs characterized by the presence of an Agent who causes the motion of the Patient (the object moved), from those motion verbs, so–called 'unaccusatives', where it seems that there is an identification between Agent and Patient (e.g.: *Clara entra las sillas* vs. *Clara anda deprisa* ('Clara brings in the chairs' vs. 'Clara walks fast')). Motion also involves the Path taken by the moved or displaced element. The path component may be defined or undefined; we can also distinguish between verbs involving a displacement and verbs involving movement without displacement (contained motion), etc. The motion class can also describe the manner, or medium, (Manner) in which the action is carried out and the part of the body involved in the movement.

Semantic components are a discrete set, or so it seems, but the (noun) objects that can flesh them out (i.e. selectional restrictions) differ according to the semantic class of verb. Hence, verbal semantic classes are defined according to the semantic components which constitute the verbal meaning, according to the possible combinations of these components and according to the possible restrictions as regards the values of the components.

### 4. 'VRQS' and lexical saturation

Once a verbal lexical meaning is decomposed into semantic components, each of them must be satisfied, or: saturated lexically, so that the meaning of that verb be completed. The saturation of these arguments can be carried out in three different ways: structurally, morphologically or by means of the so called 'understood' arguments.

a. In the first type of saturation, called structural or argumental saturation, the semantic components are saturated syntactically and independently of the verbal form. This type of saturation is related to the semantic arguments subcategorized by the verb and, therefore, it is a kind of information included in Argument Structure.

(1) *Clara* entra/sube/baja/anda/corre/salta/gatea/nalguea.
*Agent* Motion+Path+Manner+(Patient)
('*Clara goes_in/goes_up/goes_down/walks/runs/jumps/crawls or goes_on_all_fours/shakes_the_buttocks.*')

In the above example, the intransitive use of a motion verb (example 1) entails the structural or argumental saturation of the Agent component. This argument is expressed syntactically by a nominal phrase.

b. In morphological saturation, the semantic components are realized in the verbal form by means of a morphological process of lexical derivation. The expression of the morphologically saturated argument can be done either by means of prefixes and suffixes attached to the verbal form (example 2), or by means of the nominal or adjectival stem from which the verb is derived (example 3) (Baker 1988).

(2) nalgu*ear*, gat*ear*, etc. '–ear' —> manner of motion

The underlined suffix in example 2 expresses the semantic component of Manner.

In the verbs of example 3, we can observe that the argument morphologically saturated in the verbal form can be related to different semantic components: Patient, Manner, Path, etc.

(3) '*nalgu*ear' —> 'mover las *nalgas*' (Patient).
'*gat*ear' —> 'andar a *gatas*' (Manner).

c. Finally, in the last type of lexical saturation, called 'understood' saturation, the semantic components, although part of the verbal meaning, are not expressed in an explicit way, neither syntactically nor morphologically, in the verbal form. This is why we call them 'understood' arguments.

(4a) Los estudiantes bajan de forma ruidosa.
Agent Motion+Manner+Patient+*Path*
('*Students go_down noisily.*')

(4b) Los estudiantes bajan *por las escaleras* de forma ruidosa.
Agent Motion+Manner+Patient *Path*
('*Students go_down the stairs noisily.*')

This type of saturation relates to the syntactic alternation of the Path argument presented in some motion verbs (example 4). It can be structurally omitted or not.

**5. Saturation and 'diatheses'**

In this section we will focus on the syntactic consequences of the different types of lexical saturation. Semantic components can only be lexically saturated once: structurally, morphologically or in the 'understood' way. The difference between them is that the first two types of saturation can block alternations in the subcategorization patterns of verbs, while 'understood' saturation allows the alternations. So, we can establish an essential distinction between the morphological saturation carried out by a lexical derivation on one hand, and the 'understood' saturation, on the other hand.

a. In the first type of saturation, the presence of the morpheme ('*–ear*') and the noun (in italic), within the verbal form (example 5), can block the syntactic realization of the arguments represented by the suffix or the noun.

(5a) Todos *nalgu*ean cuando bailan salsa.
Agent Motion+Manner+*Patient*+Path
('*Everybody' shakes_buttocks' when they dance salsa.*')

(5b) *Todos *nalgu*ean las *nalgas* cuando bailan salsa.
Agent Motion+*Manner+Patient*+Path *Patient*
('**Everybody 'shakes_buttocks' the buttocks when they dance salsa.*')

(5c) Todos mueven las *nalgas* cuando bailan salsa.
Agent Motion+Manner+Path *Patient*
('*Everybody moves the buttocks when they dance salsa.*')

In the case of the intransitive verb *nalguear* (example 5), the morphological saturation of the Patient component in the verbal form by the presence of the noun from which the verb is derived, blocks the syntactic realization of the semantic component at issue and, therefore, any transitive alternation.[6] The only alternations allowed here are the intransitive ones: basically, the presence or absence of PP arguments. This kind of phenomenon is very extensive in the semantic field of Spanish motion verbs, in which the argument Patient is incorporated in the verbal form in order to specify the body part moved (*alear* (to flutter), *pestañear* (to blink or wink), etc.). This fact brings the possibility of distinguishing these intransitive motion verbs from those so–called unaccusatives – also intransitive – where the Patient is not a specific body part but a whole body, for which reason we say that, from a compositional point of view, there is an identification with Agent and Patient (*andar* (to walk), *correr* (to run), *bajar* (to go down), etc.).

We also can note that this sort of morphological saturation has not the same lexical consequences in the verbal argument structure. Semantic components like Agent, Patient or Path morphologically saturated have a

clear consequence in argument structure: a verb with a Patient morphologically saturated presents a monadic argument structure, while a verb with a Patient structurally saturated has a diadic argument structure. On the other hand, the morphological saturation of components like Manner does not involve, so it seems, the argument structure of that verb.

b. 'Understood' saturation finally, does not involve a lexical mark in the verbal form and does not show any blocking on the syntactic level. So, the 'understood' argument in the verbal form is like an open position that can or cannot be carried out syntactically. In this sense, this kind of saturation is richer at subcategorization alternations.

(6a) Clara baja *por las escaleras.*
Agent Motion *Path*
('*Clara goes down the stairs.*')

(6b) Clara baja ahora.
Agent Motion+*Path*
('*Clara goes_down now.*')

The semantic component Path in the motion verb *bajar* (example 6) can be represented syntactically by a prepositional phrase (example 6a) or it can be 'understood' in the verbal form (6b). When we use a motion verb like *bajar* we expect a Path so that the lexical verbal meaning may be completed, but when the argument is not syntactically expressed it can be inferred from the verbal form and, obviously, in (example 6b) the Path is less specific.

## 6. 'VRQS' in the LKB

The semantic information concerning 'VRQS' is represented by <rqs> feature in the verbal entry. Its value is always <vrqs> for the verbal signs. We have adopted this term in order to unify the information contained in lexical entries: in nominal signs we express the semantic information of nouns by the <nrqs> feature, following the initial proposal of J. Pustejovsky (1991). 'VRQS' is a complex typed feature structure in which two kinds of information are expressed: the semantic class and the way in which the semantic components of the verb are saturated.

Figure 2 shows a partial view of the semantic classes included in the 'VRQS' of the LKB's type system. The semantic components like 'medium–v', 'manner–v', 'ag–pat–v', 'path–v', etc. are specified at the first level of the hierarchy. At the successive levels, the combination of these semantic components is expressed giving as a result a semantic classification of verbs into 'motion–manner', 'motion–path', 'motion–manner–path', etc.

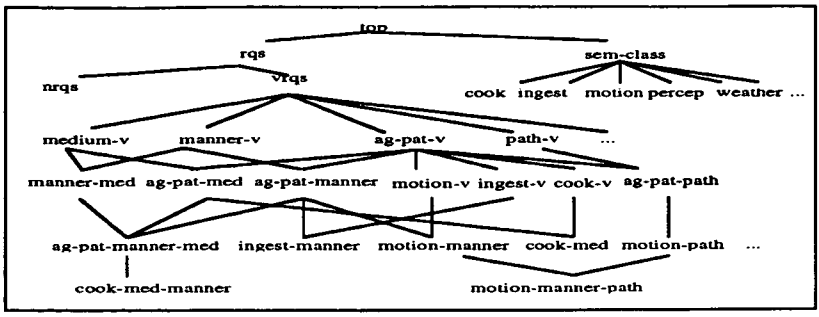


Figure 2: Type Hierarchy: Semantic components

Every semantic component has a complex structure in which the information about the type of lexical saturation presented by verbs is specified, for instance:

motion–path (ag–pat–v) Path1 (top) (OR defined undefined).
<sem–class> = motion
<morph–saturation–pth> = string
<understood–saturation–pth> = path1.

## 7. Conclusions

We have presented a first approach to 'VRQS' representation in a Lexical Knowledge Base, which describes the verbal lexical meaning in terms of its semantic components. We have shown how 'VRQS' could encode some phenomena of saturation and how this information has important syntactic consequences: we can block a specific subcategorization alternation according to the type of argument saturation.

So far, we have worked with six different semantic classes of verbs: cooking, communication, ingestion, motion, perception and weather verbs. The future aim of the research is to develop more semantic classes in order to have enough complete representation for different verbs.

### Notes

1 The research reported here was undertaken in the framework of the Acquilex–II Project (Esprit 7315) "The Acquisition of Lexical Knowledge for Natural Language Processing Systems" at the Universitat Politècnica de Catalunya.

2 Copestake (1992) shows a complete treatment of the functionality of the LKB and its representation language (LRL). Other interesting documents about the LKB are Ageno et al. (1992), Copestake et al. (1991), Sanfilippo (1991).

3 We will not discuss the information referring to argument structure, event structure and selectional restrictions as they are adequately treated in Sanfilippo (1991).

4 The feature 'sense–id' specifies the information about the source dictionary, language, sense, etc. of the entry.

5 Sanfilippo (1991)
6 "Generación de Alternancias de Subcategorización via Reglas Léxicas" (Taulé, M. et Castellón, I. 1993) presents a treatment of the transitives subcategorization alternations in Spanish, and establishes a classification for them.

**References**

Acquilex–II, 1992. *The Acquisition of Lexical Knowledge* .Basic Research Project 7315 ACQUILEX–II, Technical Annexe.

Ageno, A., I. Castellón, M.A. Martí, F. Ribas, G. Rigau, H. Rodríguez, M. Taulé, M.F. Verdejo, 1992. "SEISD: An Environment for Extraction of Semantic Information from On–line Dictionaries". *Proceedings of the 3rd Conference on Applied Natural Language Processing.* Trento.

Baker, M.1988. *Incorporation: A theory of Grammatical Function Changing.* Chicago: University of Chicago Press.

Copestake, A. 1992."The ACQUILEX LKB: Representation Issues in Semi–Automatic Acquisition of Large Lexicons". *Proceedings of the 3rd Conference on Applied Natural* Language Processing. Trento.

Jackendoff, R. 1990. *Semantic Structures.* Cambridge, MA.: MIT Press.

Levin, B.1993. *English Verb Classes and Alternations.* Chicago:University of Chicago Press.

Pollard, C. & Sag, I.1987. *An Information–Based Approach to Syntax and Semantics: Fundamentals.* Vol. I. Stanford: CSLI.

Pollard, C. & Sag, I.1992. *Agreement, Binding and Control: Information–Based Syntax and Semantics.* Vol.2. Stanford: CSLI.

Pustejovsky, J.1991. "Generative Lexicon" in *Computational Linguistics.*, 17.

Sanfilippo, A. 1991. "LKB Encoding of Lexical Knowledge". Briscoe, T., Copestake, A., & de Paiva, V. (eds.).

Sanfilippo, A.1992. "Verbal Diatheses: Knowledge Acquisition, Lexicon Construction and Dictionary Compilation" (forthcoming).

Talmy, L.,1985. 'Lexicalization patterns" in Shopen, T. (ed) *Language Typology and Syntactic Description* vol. 3 (Grammatical Categories and the Lexicon). Cambridge: CUP.

Taulé, M. & Castellón, I. 1993. "Generación de Alternancias de Subcategorización via Reglas Léxicas" *Actas del IX Congreso anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)* Santiago de Compostela: SEPLN.