In I. Moskowich-Spiegel et al. (eds.) (2010). ***Language Windowing through Corpora*****, pp. 207-214. ISBN: 978-84-9749-401-4**

A corpus based approach on event structure: simple and complex predicates of Spanish

MARTA COLL-FLORIT

Universitat Oberta de Catalunya

JUAN APARICIO

IRENE CASTELLÓN

Universitat de Barcelona

In order to modelize the behavior of event structure for the simple and complex verbal predicates of Spanish, we use corpora to obtain empirical evidence of verbal behavior. We assume the hypothesis that the Aktionsart *is compositional. However, unlike the traditional approach, we start from the basic hypothesis that the verbal units have different lexical weights or degrees of interaction between the lexicon and the grammatical construction. The data show that aspectual categories impose different contextual restrictions and that there exist different degrees of prototypicity within each category –some members are more stable / flexible than others-. The gradual internal structuring of each category is not arbitrary; it is due to the semantic characteristics shared by the different aspectual verbal subsets.*

Keywords: Aktionsart, *prototypicity, corpus linguistics*

Para poder modelizar el comportamiento eventivo de los predicados verbales simples y complejos del español, utilizamos corpus para obtener evidencia empírica del comportamiento verbal. Asumimos la hipótesis de que el Aktionsart *es composicional, sin embargo, no como en la aproximación tradicional, nuestra hipótesis de partida básica es que las unidades verbales presentan diferentes pesos léxicos o grados de interacción entre el léxico y la construcción gramatical. Los datos muestran que las categorías aspectuales imponen diferentes restricciones contextuales y que existen diferentes grados de prototipicidad dentro de cada categoría –algunos miembros son más estables / flexibles que otros-. La estructura gradual interna de cada categoría no es arbitraria, sino que se debe a las características semánticas que comparten los diferentes subconjuntos verbales aspectuales.*

Palabras claves: Aktionsart, *prototipicidad, lingüística de corpus*

1. INTRODUCTION

The main goal of this paper is to present the research[1] that we have carried out in order to modelize a representation of event structure for the simple and complex predicates of Spanish. In the different existing computational resources (*FrameNet* (Subirats, 2005), *Adesse* (García-Miguel, Costas and Martinez, 2005), *AnCora* (Aparicio, Taulé and Martí, 2008), *WordNet* (Miller, Beckwith, Fellbaum, Gross and Miller, 1990)), either this

[1] This research has been carried out within the KNOW2 project granted by Ministry of Education and Science, Spain Government (Semantic Knowledge Representation (SKR), TIN2009-14715-C0403).

information is not present or the analysis tends to be quite simplified. Moreover, periphrases are not represented at all or, if they are, the representation is not made from an event structure perspective.

From a methodological point of view, it is difficult to find aspectual studies that contain a significant amount of predicates; generally, the analysis is reduced to a set of predicates (Dowty, 1979; Verkuyl, 1993; among others). Accordingly, our interest is to work with a wide and varied number of predicates in order to obtain a representation system based on empirical methods. With this idea in mind, we use corpora to obtain empirical evidence of verbal behavior, which is the main goal of this paper. In addition to that, we carry out psycholinguistic experiments to determine the basic aspectual properties that human beings distinguish in language processing (Coll-Florit and Gennari, 2009). These experiments will allow us to use these properties as basic elements in the predicates representation, as well as to validate the data obtained from corpora.

Our starting points are the four ontological types of events (states, activities, accomplishments and achievements), which correspond with the classification of Vendler (1967) and Dowty (1979). We also assume the hypothesis that the *Aktionsart* is compositional. However, unlike the traditional approach (Verkuyl, 1993), we start from the basic hypothesis that the verbal units have different lexical weights or degrees of interaction between the lexicon and the grammatical construction. In particular, depending on the degree of aspectual flexibility that the verbs accept, we assume that three types of predicates can be identified, following Coll-Florit (2009): stable monosemic verbs (e.g. *equivaler* ‘to be equivalent’), flexible monosemic verbs (e.g. *perder* ‘to lose’), and polysemic verbs (e.g. *contener* ‘to include / to hold back’), a typology that is reflected in different degrees of syntactic, semantic and morphological stability -from less to more flexible, respectively-.

Another theoretical difficulty that we are now analyzing refers to verbs participating in complex predication and their effects on the event structure. Aspectual periphrases are sensitive to and may produce effects on the *Aktionsart* of the predication (Dick, 1989). For example, the Egressive Aspect, such as the one expressed by means of *terminar de* ‘to finish’ can only modify dynamic predications and it renders the predication telic. Therefore, the context can provoke that some verbal predicates may move towards other aspectual classes.

2. STUDY OF CORPORA

We focused on monosemic verbs. The study was reduced to two basic aspectual categories: states (non-dynamic and durative events) and achievements (dynamic and punctual events). The aims of this study were:

1. To test whether these aspectual categories imposed different contextual restrictions;
2. To check out if there were different degrees of prototypicity within each category -some members are more stable / flexible than others-, which may imply movement to another aspectual categories.

The procedure of the study was divided into two phases. In the first one, a total of 60 simple Spanish predicates were analyzed: 30 belonging to states and 30 to achievements. We have considered a total of 14 grammatical constructions that, in the bibliography on *Aktionsart*, are used in a regular way to identify parameters such as dynamism, delimitation and / or duration. In the second phase, in order to confirm the results, we extended the study to exactly the same group of verbs when appearing as complex predicates, specifically as aspectual periphrases.

The basic methodology consisted on analysing the frequencies of appearance of each verb for the total number of constructions. In particular, with reference to simple predicates, the study was based on a subcorpus of 81 million words from the *Corpus de Referencia del Español Actual* by the *Real Academia Española*[2]. As for complex predicates, the study was based on a subcorpus of 23 million words from the *Corpus del Español*[3].

The procedure of the analysis was divided into two stages. In the first one, an *intercategorial* analysis was carried out in order to check out if the two selected aspectual categories (states and achievements) presented different patterns of use. In the second stage, an *intracategorial* analysis was carried out so as to identify which was the set of verbs of a category that appeared with the highest and lowest frequency for each construction.

The results of the intercategorial analysis clearly show different patterns of morphosyntactic use for states and achievements. For instance, a clear interaction has been noticed between the lexical aspectual type and the frequency of appearance with a determined verbal tense: stative predicates, inherently non-delimited, are more frequent with imperfective tenses, whereas predicates that express achievements, inherently delimited, present the highest frequency rate with perfective tenses (table 1).

[2] http://www.rae.es

[3] http://www.corpusdelespanol.org

Table 1: Intercategorial analysis of the simple predicates: verbal tense

	Present	Imperfect	Past Simple
States	40 %	15%	3,6%
Achievements	16 %	4%	16%

Moreover, the results show equivalent patterns of frequency among constructions that imply the same aspectual parameter, as well as inverse patterns for those contexts that imply opposite parameters (table 2).

Table 2: Intercategorial analysis of the simple predicates: the durative parameter

		States	Achievements
Punctual constructions	*De repente* 'Suddenly'	0,003%	0,14%
	A las X horas 'At X'	0,0001%	0,07%
Durative constructions	*Desde hace X tiempo* 'X time ago'	0,4%	0,009%
	Durante X tiempo 'For X time'	0,7%	0,05%

According to the results of the intracategorial analysis, aspectual categories show a gradual internal structuring, which is not arbitrary; it is due to the semantic characteristics shared by the different aspectual verbal subsets.

Within the states, three subtypes differing in aspectual flexibility have been identified: permanent, transitory and psychological states. The predicates that express permanent states (e.g. *equivaler* 'to be equivalent', *caber* 'to fit') are the most prototypical of the category, since they do not accept either dynamic or punctual constructions. Regarding the transitory states (e.g. *estar preocupado* 'to be worried', *estar enfermo* 'to be ill'), they occupy an intermediate position since they admit constructions that delimit the temporary period in which the situation takes place. Finally, the psychological states (e.g. *conocer* 'to know', *creer* 'to believe', *gustar* 'to like') are the most flexible ones. Specifically, they are closer to achievements when they appear in the simple past: *Joaquín, de 25 años, conoció a la empresaria* [‘Joaquín, 25 years old, met the businesswoman’], or in the immediate prospective periphrasis: e*stán a punto de conocerse* [‘They are about to know each other’]. However, when they appear with an ingressive periphrasis, they are closer to processes: *a Oliveira le empezó a gustar más el cigarillo* [‘Oliveira began to like more the cigarette’].

Regarding achievements, they also present an internal gradation. On the one hand, there are prototypical punctual predicates that do not accept durative constructions (e.g. *atrapar*, ‘to catch’, *detectar*, ‘to detect’) and, on the other hand, there are more flexible predicates that accept changes of aspectual interpretation (e.g. *perder*, ‘to lose’, *cerrar*, ‘to close’). More precisely, this last verbal group accepts durative constructions focusing on the resulting state of the achievement: *cerraron las instalaciones durante una semana* [‘They closed the facilities during a week’]. When this happens, they are closer to states. This subset of verbs also accepts ingressive periphrases: *comenzó a perder la vista a los diez años* [‘He began to lose his sight when he was ten years old’]. In this case, they profile the ingression of a process. Finally, this group of verbs also admits egressive periphrases: *ya ha terminado de cerrar la puerta,* [‘She has already finished closing the door’]. In this case, they are closer to accomplishments since they focus on a delimited durative event.

We have just seen how verbal predicates may move towards other aspectual classes, with unequal different lexical strength -some elements are more prototypical than others-.

3. TOWARDS A REPRESENTATION OF EVENT STRUCTURE

Several resources such as *WordNet*, *SenSem* or *AnCora* include event structure information in the semantic characterization of verbal predicates. However, these lexical resources present three basic problems. First of all, each resource adopts a different typology of classification, which makes it difficult to establish equivalence relations between them. Secondly, neither the movement between aspectual categories, nor the different degrees of prototypicity within each category are represented. Finally, and in a related way, the different existing lexical resources codify the verbal senses ‘per se’, without considering the emergence of periphrastic combination units.

Given these lacks, and considering the results of the present study, we understand that the computational representation of the event structure has to include, at least, the interaction of three basic phenomena for each verbal sense:

1. Inherent event structure configuration. This factor will have to reflect the association prototypicity of the predicates with the different classes;
2. Description of the incompatible constructions and/or constructions that imply a change of aspectual interpretation (where the verbal periphrases would be included);
3. Semantic type of the complements, a key factor when representing aspectual polysemic verbs, as long as each sense imposes different selection restrictions (Coll-Florit, 2009).

4. CONCLUSIONS AND FUTURE WORK

We have presented the research that we have carried out in order to reach, in a near future, a representation of event structure based on empirical methodology. We have also presented the theoretical problems we have come across and the solutions that we propose. According to this framework we have presented empirical studies based on corpora and the results obtained for simple as well as for complex predicates. The data show that aspectual categories impose different contextual restrictions and that there exist different degrees of prototypicity within each category. Finally, we have pointed out the main characteristics that our representation system is going to have.

For the future work, it is necessary to describe this representation system in detail, to decide the formal language that will be used, and to define in which way the properties associated with the different classes and with the predicates will be implemented. This representation system must reflect the degree of prototypicity of a predicate in relation to its class. Moreover, it must be related to the capacity of a predicate to move to other aspectual classes depending on the context.

REFERENCES

Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (ed.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*. John Benjamins Publishing Co.

Aparicio, J. M. Taulé, M.A. Martí (2008). “Ancora-Verb: A Lexical Resource for the Semantic Annotation of Corpora”, *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh.

Coll-Florit, M. (2009). *La modalitat de l'acció. Anàlisi empírica, reformulació teòrica i representació computacional*. PhD Dissertation. IN3/UOC.

Coll-Florit, M., S. Gennari (2009). "Time in language: event duration in language comprehension", *Proceedings of the 22nd Annual* CUNY *Conference on Human Sentence Processing*. University of California.

Dick, S.C. (1989). *The Theory of Functional Grammar*. *Part 1*.: The structure of the Clause. Dordrecht: Foris.

Dowty, D. (1979). *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Dordrecht: Reidel.

García-Miguel, J. M., L. Costas, S. Martínez (2005): "Diatesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE", Wotjak, Gerd, & Juan Cuartero Otal (eds.), *Entre semántica léxica, teoría del léxico y sintaxis*. Frankfurt am Main: Peter Lang.

Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K. Miller (1990). *Five Papers on Wordnet*. CSL Report 43. Cognitive Science Laboratory, Princeton University.

Subirats Rüggeberg, C. (2005). "FrameNet español. Una red semántica de marcos conceptuales". E. Serra, G. Wotjak, eds. *Cognición y percepción lingüísticas*. Valencia: Universidad de Valencia and Universidad de Leipzig, pp. 182-196.

Vendler, Z. (1967). *Linguistics in Philosophy*. Ithaca, N.Y.: Cornell University Press.

Verkuyl, H.J.(1993). *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure*. Cambridge: Cambridge University Press.