

---

Cristina Sánchez-Marco, Stefan Evert  
Universitat Pompeu Fabra, Barcelona, Institute of Cognitive Science, University of  
Osnabrück  
[cristina.sanchezm@upf.edu](mailto:cristina.sanchezm@upf.edu), [stefan.evert@uos.de](mailto:stefan.evert@uos.de)

## **Measuring semantic change: The case of Spanish participial constructions**

### **An empirical study of semantic change**

The goal of this paper is to explore different quantitative measures of semantic change and thus contribute to the development of statistical techniques for diachronic studies of meaning. Specifically, we focus on semantic change in the Spanish participial constructions *haber* ‘have’ + participle, *ser* ‘be’ + participle, *estar* ‘be/stay’ + participle, *tener* ‘have/possess’ + participle.

In 20th-century Spanish these constructions have clearly distinct meanings: they express the perfect, the verbal passive, the resultant state passive and a stative meaning, respectively. In contrast, the meanings of these participial constructions before the 17th century were quite different. *Ser* + participle was mainly used as a passive, but could also be a perfect for some intransitive and reflexive verbs, as well as a resultant state passive. *Haber* + participle could also have a stative meaning similar to the construction *tener* + participle in contemporary Spanish. *Estar* + participle and *tener* + participle emerged for the first time in the 13th century, but it took them more than 300 years to

generalise their meanings as they are today. The following examples illustrate the use of *ser* + participle as a passive nowadays and its use as a perfect in the earlier centuries.

- (1) *Su monotonía es interrumpida solamente por algunas llamadas quejumbrosas.*  
His monotony is interrupt<sub>PTCP.F.S</sub> only by some calls plaintive  
'His monotony is interrupted only by a few plaintive calls.' (20th c.)
- (2) *Myo Çid Ruy Diaz a Alcolçer es venido.*  
Myo Çid Ruy Diaz to Alcolçer is come<sub>PTCP.M.S</sub>  
'Myo Çid Ruy Diaz has come to Alcolçer.' (12th c.)

## Approach

There is a large body of literature on semantic change, proposing different explanations for the change in Spanish participial constructions (Mendeldoff 1964, Vincent 1982, Pountain 1985, Aranovich 2003, Copple 2009). However, we are not aware of any empirical study that supports these claims with quantitative evidence. Our goal is to fill this gap, using data from a representative diachronic corpus of Spanish (see the section Data below).

The empirical study reported here addresses two basic questions:

- A. Is there a *significant change* in the frequency and usage of all four participial constructions?
- B. If a significant change took place, *how* and *why* did it take place?

We argue that quantitative data can play an important role in answering such questions. In particular, we explore the following approaches in order to measure and explain semantic change:

1. *Frequency and productivity of participial constructions.* We track changes in the frequency and productivity of all four constructions from the 12th to the 20th century. Our hypothesis is that *specialization* of *ser* + participle as the passive leads to a decrease both in frequency and productivity, as the auxiliary is applicable in fewer contexts and can be combined with fewer types of predicates. Conversely, *grammaticalization* of *haber* + participle as a perfect should lead to an increase in frequency and productivity.
2. *Distributional measures of semantic variability.* Following Sagi et al. (2009), we define the semantic density of a construction as the average distance between distributional representations of its instances in the corpus. If the construction undergoes specialization, it should be applicable in a narrower range of contexts and hence its semantic density should increase (*mutatis mutandum* for grammaticalization). Expanding on this token-based method, we also compare distributional representations of the types generated by the construction with those of the corresponding base predicates.
3. *Similarity as an explanation for semantic change.* We use distributional methods to determine the semantic similarity between the auxiliaries *estar*, *ser*, *tener* and *haber* in different centuries. Our hypothesis is that semantic similarity of *estar* with *ser* and of *tener* with *haber* might have contributed to the emergence of the

## Data

The data for this study have been retrieved from a Spanish diachronic corpus consisting of 651 documents from the 12th to the 20th century, with a total size of more than 40 million words. The corpus comprises a variety of genres (fiction and nonfiction) and the documents come from different sources: Data from the 12th century to the 1950s were collected from the electronic texts transcribed and compiled by the *Hispanic Seminary of Medieval Studies*<sup>1</sup>, the *Gutenberg project*<sup>2</sup> and the *Biblioteca Cervantes*<sup>3</sup>. This part of the corpus has been annotated automatically with linguistic information (morphosyntactic tag and lemma), using an expanded version of the Freeling morphological analyzer (Sánchez-Marco et al. 2010). Tagging accuracy in the oldest texts of the corpus yields 92%, which is sufficient to make reliable statistical generalizations.<sup>4</sup> Additional texts from the years 1975 to 1995 were obtained from the *Lexesp corpus* (Sebastian-Gallés 2000).

In order to facilitate statistical analysis of the data, we divided the corpus into four main periods, following the customary division determined by external historical events: Middle Spanish (1100-1492), Modern Spanish (1493-1788), Contemporary Spanish (1789-1974), and Late Contemporary Spanish (1975-2000).

## Results

So far, we have completed a detailed study of changes in the usage frequency of each participial construction from the 12th to the 20th century. Frequency counts were obtained using the IMS Open Corpus Workbench<sup>5</sup> and analyzed with the open-source statistical software R (R Development Core Team 2010).

Figure 1 illustrates the development of *haber* + participle (left column) and *ser* + participle (right column). Each point in the two bottom panels corresponds to a single text from the corpus, showing time of composition on the x-axis and the relative frequency of the corresponding construction on the y-axis. From these graphs, it is obvious that the frequency of *haber* + participle increases continuously from the 12th to the 20th century; the change accelerates from the 15th century on. The frequency of *ser* + participle decreases, especially during the 15th and 16th centuries.

The boxplots in the top row of Figure 1 compare pooled data for the four main periods. All differences between Middle, Modern and Contemporary Spanish are highly significant (Generalized Linear Model with binomial family and logit link,  $p < .001$ ). We interpret these findings as evidence for the grammaticalization of *haber* + participle and the specialization of *ser* + participle.

---

<sup>1</sup> See Corfis et al. [1997], Herrera and de Fauve [1997], Kasten et al. [1997], Nitti and Kasten [1997], O'Neill [1999], Sánchez et al. [2003].

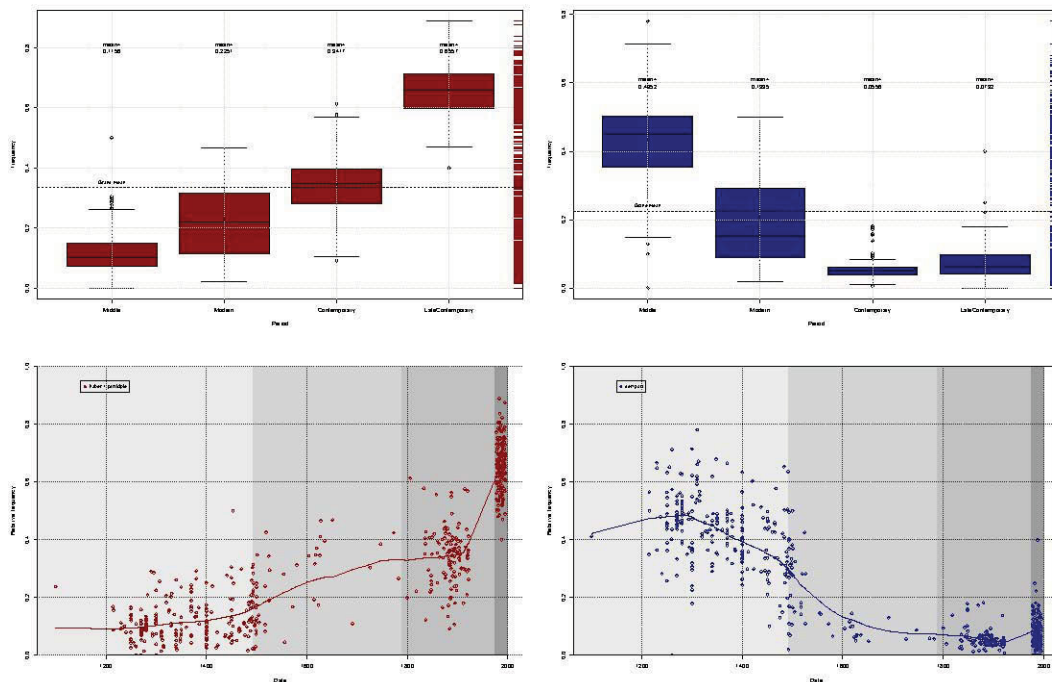
<sup>2</sup> [http://www.gutenberg.org/wiki/Main\\_Page](http://www.gutenberg.org/wiki/Main_Page).

<sup>3</sup> <http://www.cervantesvirtual.com/>.

<sup>4</sup> Sánchez-Marco, p. c.

<sup>5</sup> <http://cwb.sourceforge.net/>.

FIGURE 1. FREQUENCIES OF *HABER* + PARTICIPLE AND *SER* + PARTICIPLE FROM THE 12TH TO THE 20TH CENTURY.



## Work in progress

We are currently working on the other quantitative approaches described in the section Approach above. In particular, we have compiled distributional semantic models for each century, which allow us to derive (i) *type vectors* representing each type generated by a specific construction in a particular century and (ii) *context vectors* representing token instances of each construction in a particular century. Based on these data sets, we will perform experiments with methods 2 and 3.

We are also collecting type-token statistics for the participial constructions. These data will be analysed with LNRE models of quantitative productivity (Baayen 2001) in order to complete our application of method 1. We intend to use the open-source R package *zipfR* (Evert & Baroni 2007) for this purpose.

## References

- Aranovich, Raúl 2003. The semantics of auxiliary selection in Old Spanish. *Studies in Language*, 27:1-37.
- Baayen, R. Harald 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.
- Copple, Mary T. 2009. *A Diachronic Study of the Spanish Perfect(ive): Tracking the Constraints on a Grammaticalizing Construction*. PhD thesis, University of New Mexico.
- Corfis, Ivy A., O'Neill, John and Beardsley, Theodore S. Jr. (eds.) 1997. *Early Celestina Electronic Texts and Concordances*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.
- Evert, Stefan and Baroni, Marco 2007. *zipfR*: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29-32, Prague, Czech Republic, June.
- Herrera, María Teresa and González de Fauve, María Estela (eds.) 1997. *Concordancias electrónicas del corpus médicos español*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.

- Kasten, Llyod, Nitti, John and Jonxis-Henkemans, Wilhemina (eds.) 1997. *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.
- Mendeldoff, Henry 1964. The passive voice in old spanish. *Romanistisches Jahrbuch*, 15:269-287.
- Nitti, John and Kasten, Llyod (eds.) 1997. *The Electronic Texts and Concordances of Medieval Navarro-Aragonese Manuscripts*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.
- O'Neill, John 1999. *Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.
- Pountain, Christopher 1985. Copulas, verbs of possession and auxiliaries in Old Spanish: The evidence for structurally interdependent changes. *Bulletin of Hispanic Studies*, 62:337-355.
- R Development Core Team 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Sagi, Eyal, Kaufmann, Stefan and Clark, Brady 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens, Greece.
- Sebastián-Gallés, Núria 2000. *LEXESP: léxico informatizado del español*. Edicions Universitat Barcelona.
- Nieves Sánchez, Maria, Herrera, María Teresa and Zabía, María Purificación 2003. *Textos medievales misceláneos*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.
- Sánchez-Marco, Cristina, Boleda, Gemma, Fontana, Josep Maria and Domingo, Judith 2010. Annotation and representation of a diachronic corpus of Spanish. In *Proceedings of LREC*, Valletta, Malta. Online at [http://www.lrec-conf.org/proceedings/lrec2010/pdf/535\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/535_Paper.pdf) [March 05, 2011].
- Vincent, Nigel 1982. The development of auxiliaries *habere* and *esse* in Romance. In *Studies in the Romance Verb*, Vincent, Nigel and Harris, Martin (eds.), 71-96. London: Croom Helm.
-