Measuring lexical extension: The case of Spanish *estar* + past participle

Cristina Sánchez-Marco[1], Rafael Marín[2], Stefan Evert[3]
[1]Universitat Pompeu Fabra, Barcelona,[2]CNRS - Université Lille, [3]Technische Universität Darmstadt
[1]cristina.sanchezm@upf.edu, [2]rafael.marin@univ-lille3.fr,
[3]evert@linglit.tu-darmstadt.de

1 Introduction

In this paper we present a novel approach to studying language change through (lexical) extension. This approach draws both on current linguistic theories and on the quantitative analysis of corpus data in order to achieve a deeper understanding of the nature of extension in language change.

As a case study, we explore change in *estar* 'to stay' + past participle (PTCP) from the 13th to the 20th century in a diachronic corpus of Spanish. Specifically, we claim that the extension of this participial construction takes place gradually through the lexicon, in that it spreads from some types of predicates to others and it is driven by certain semantic restrictions. Furthermore, we argue for the idea that this extension is mainly motivated by the competition between *estar* + PTCP and *ser* 'to be' + PTCP (for some of their readings).

The contents of this paper are as follows. In §2 we illustrate with examples the uses and interpretations of *estar* + PTCP and *ser* + PTCP in Old and Modern Spanish and briefly summarise how change in these constructions has been analysed in previous studies. In §3 we argue for the idea that change in *estar* + PTCP is mainly motivated by the competition against *ser* + PTCP for the result state reading and that it is driven by general mechanisms of language change. We then move to §4, where we present the approach and data used to argue for these ideas, and §5, where we provide quantitative evidence. Finally, we conclude the paper in §6 with some discussion and further work.

2 The problem

If one observes some occurrences of *estar* + PTCP in a diachronic corpus, one soon notices that *estar* + PTCP has substantially changed its readings from Old to Modern Spanish. In Modern Spanish, *estar* + PTCP is used to describe the result state of a process involving a particular entity denoted by the subject of the construction; whereas in Old Spanish this construction was interpreted either as a result state, an eventive passive, or even as a perfect.

Precisely two of the possible readings of *estar* + PTCP in Old Spanish (result state and eventive passive) were also conveyed by another participial construction in the same

stages of the language, specifically *ser* + PTCP, which in time has become the only means to express eventive passives in Modern Spanish. In general terms, change in these constructions can thus be seen as an specialization of *estar* + PTCP and *ser* + PTCP to express result states and eventive passives, respectively.[1]

(1) and (2) show some 13th century examples of result states expressed using either *estar* or *ser*. In (3) and (4) there are some examples that illustrate uses of these constructions in the 20th century.

(1) *E yo commo estaua desesperado & me enojaua ya de beujr eneste*
and I as stayed despaired and me got-angry already of live in-this
mundo.
world.
'And as I was desperate, I got fed up with living in this world.' (Result state *estar*+PTCP, 13th c.)

(2) *Fueron todos desesperados de meiorar en su fazienda.*
were all despaired of improve in their possessions
'All of them were desperate to increase their possessions.' (Result state *ser* + PTCP, 13th c.)

(3) *Los equipajes estaban amontonados en la sala de baile.*
the luggages stayed piled in the room of dancing
'The luggages lay piled up in the dancing room.' (Result state *estar* + PTCP, 20th c.)

(4) *Este asunto fue provocado por un acuerdo secreto entre el Jefe de*
this matter was prompted by a agreement secret between the head of
Gobierno y tú.
government and you
'This matter was prompted by a secret agreement between you and the head of government.' (Eventive passive *ser* + PTCP, 20th c.)

The main question we aim to answer in this paper is how *estar* + PTCP was established as the only means to express result states. In this paper we argue for the idea that mainly two factors have driven this change: extension as a mechanism and competition against *ser* + PTCP for the same readings as a motivating force.

Both the extension of *estar* + PTCP as the only means to express result states and the relation between the development of the two constructions *estar* + PTCP and *ser* + PTCP has often been noted by previous research (Bouzet, 1953; Mendeloff, 1964; Pountain, 1985; Yllera, 1980; Batllori and Roca, 2011). However, most of these studies focus on describing either particular stages in the development of *estar* + PTCP or on

[1]The focus in this paper is only on the extension of *estar* + PTCP to express result states and therefore we leave the study of change in instances of perfect *estar* + PTCP for future research.

investigating the reanalysis of the grammatical structure of the construction from the source to the target interpretation. For example, Pountain (1985) suggests that the result state interpretation of *estar* + PTCP increases steadily at the expense of *ser* + PTCP at least until the 16th century. However, apart from the fact that these authors base their conclusions on very small corpora; more importantly, they provide no clear account for the motivation behind the gradual change in the rates these constructions are used with the given meanings across time.

To our knowledge there is no study that addresses the question of how exactly the spread of *estar* + PTCP as a result state occurs, specifically which factors, constraints and mechanisms may have driven this change. In this paper we try to inquire into these questions and offer some answers for them validated on corpus data.

3 Hypotheses

The main hypothesis to be argued in this paper is that extension of *estar* + PTCP to express result states takes place gradually through the lexicon and that it is mainly motivated by the competition against *ser* + PTCP for the same readings.

Extension is generally considered a change in the surface expression of a linguistic item or construction. For example, Harris and Campbell (1995) define extension as the generalization of an already existing rule in grammar, such as the extension of a set of case marking rules at the expense of another more complex set of rules in Laz, a language from the Kartvelian family. Lexical extension (also called lexical diffusion) is the term used to refer to a type of extension which happens gradually through the lexicon, one or several words at time. In the literature this notion has often been used to explain how sound change happens in language, in that it is phonetically abrupt but lexically gradual or spreading word by word through the lexicon (Labov, 1981). On the other hand, lexical extension has also been used to explain how some semantic and syntactic changes extend in time: in a gradual way and through the lexicon. In lexical extension each context constitutes a new locus for change. For example, Joseph (1983) studies lexical extension of the loss of the infinitive in Greek and other Balkan languages. More recently, Rodríguez-Molina (2004) has followed a similar approach in explaining the development in the Spanish perfect *haber* + PTCP. In some approaches to syntactic change, the extension of change through the lexicon has been modelled using a logistic function, which underlies the largely attested S-shaped curve of linguitic change (Kroch, 1989).

It is often the case that extension of a certain form takes place as the result of competition between two or more forms. Specifically, competition usually takes place when two or more forms are alternative expressions for the same context of use or interpretation.When two forms appear in the same contexts and have the same interpretations some kind of competition eventually happens, which often leads to the replacement of one at the expense of the other (as a consequence of what may be considered some kind of *blocking effect*, in the sense of Aronoff (1976)). Competition then leads to an

eventual replacement of one of the forms. This replacement takes place in a gradual way. Speakers will increasingly select one form over the other. Consequently, the selected form conventionalizes its use in language and ultimately replaces the non-selected form, which disappears as an alternative form.

In this paper we argue that precisely this type of change is what happens to *estar* + PTCP, namely a lexical extension of the construction as the only means to express result states which is mainly motivated by the competition against *ser* + PTCP for the same interpretation. More specifically, the hypothesis proposed in this paper is that *estar* + PTCP spreads gradually through the lexicon from certain types of verbs to others (we refer to it as the *Lexical Extension Hypothesis*).

The contexts of principal relevance for such an extension are those in which *estar* is combined with psychological verbs having an object experiencer, such as *preocupar* 'worry'. To account for this change, we argue that *preocupar* psychological verbs fit well into the syntax and semantics of *estar*, which in its origins was used to describe the position of an entity. Based on Gawron's (2005) analysis of the semantics of extent predicates and Sánchez-Marco and Marín's (2011) proposal that the same semantic component that exists in extent predicates is also part of the denotation of object experiencer psychological verbs, we assume that these verbs encode an abstract locative component (or *path*) in their meaning, which enable them to fill the position of the path argument of *estar*. As a second step, the construction extended to other types of predicates, specifically to accomplishments and achievements.

4 Approach & Data

To go one step further in the use of corpus evidence to explore extension in language change, in this paper we propose an approach that combines current linguistic theories, in particular formal semantics (e.g. Dowty, 1979; Gawron, 2005, 2009; Beavers, under review; Zwarts, 2006), with quantitative methods, such as Generalized Linear Models (GLMs, see Baayen, 2008). This approach allows us to keep track of the specific conditions, constraints and motivations which are relevant for extension, and at the same time preserve the basic fact that change takes place over a long period of time and in a gradual way through generations of speakers. The use of GLMs guarantees a sound statistical analysis of corpus data.

The data used for the case study presented in this paper have been retrieved from a diachronic corpus of Spanish, containing more than 40 million words from the 13th to the 20th century and comprising a wide variety of genres and styles. The documents forming this corpus come from different sources: Data from the 13th century to the 1950s were collected from the electronic texts transcribed and compiled by the *Hispanic Seminary of Medieval Studies*[2], the *Gutenberg project*[3] and the *Biblioteca*

[2]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).

[3]`http://www.gutenberg.org/wiki/Main_Page`

Cervantes[4]. This part of the corpus has been annotated automatically with linguistic information (morphosyntactic tag and lemma), using an extended version of the Freeling morphological analyzer (Sánchez-Marco *et al.*, 2011). Additional texts from the years 1975 to 1995 were obtained from the *Lexesp corpus* (Sebastián-Gallés, 2000).

In order to explore the lexical extension of *estar* + PTCP through verb types, lists of verbs from 4 different semantic classes are used: accomplishments (50 verbs; e.g. *arrasar* 'to devastate', *arreglar* 'to fix', *ascender* 'to ascend', ...), achievements (50 verbs; e.g. *abrir* 'to open', *acertar* 'to guess', *adquirir* 'to acquire', ...), object experiencer psych verbs (100 verbs; e.g. *agobiar* 'to overwhelm', *animar* 'to encourage', *motivar* 'to motivate', ...), and subject experiencer psych verbs (25 verbs; e.g *admirar* 'to admire', *adorar* 'to adore', *amar* 'to love', ...). These lists have been carefully compiled by the authors based on standard linguistic diagnostics that are sensitive to certain semantic features, such as telicity or homegeneity.

Frequency counts of all combinations of *ser* and *estar* with past participles of these verbs in the corpus were obtained using the IMS Open Corpus Workbench[5] and analyzed with the open-source statistical software R (R Development Core Team, 2010).

5 Evidence

So far, we have completed a detailed study of changes in the usage frequency of each participial construction from the 13th to the 20th century. Preliminary results of this analysis can be seen in figure 1. Each point in the two top figures corresponds to a single text from the corpus, showing time of composition on the x-axis and the relative frequency of the respective construction on the y-axis. From these graphs, it is obvious that the frequency of *estar* + PTCP increases continuously from the 13th to the 20th century, whereas the frequency of *ser* + PTCP decreases dramatically, especially during the 15th and 16th centuries. We interpret the results of this study as evidence for lexical extension of *estar* + PTCP and the competition against *ser* + PTCP. All frequency differences between historical periods (indicated by different shades of grey in the plots) are highly significant (Generalized Linear Model with binomial family and logit link, $p < .001$).

Figure 2 represents the relative percentages of *estar* (in green) and *ser* (in blue) with past participles of object experiencer psychological verbs. From the graph it is obvious that the frequency of *estar* with *preocupar* verbs in comparison with the frequency of *ser* with the same type of verbs increases continuously from the 13th century and quite dramatically from the 16th to the 20th century. We take this as evidence for the lexical extension of *estar* + PTCP as it was explained before.

We are currently working on statistical significance tests for changes in the frequencies of *ser* and *estar* with participles of other types of psychological verbs, accomplishments and achievements. Preliminary results of this analysis show that the increase in

[4] `http://www.cervantesvirtual.com/`

[5] `http://cwb.sourceforge.net/`

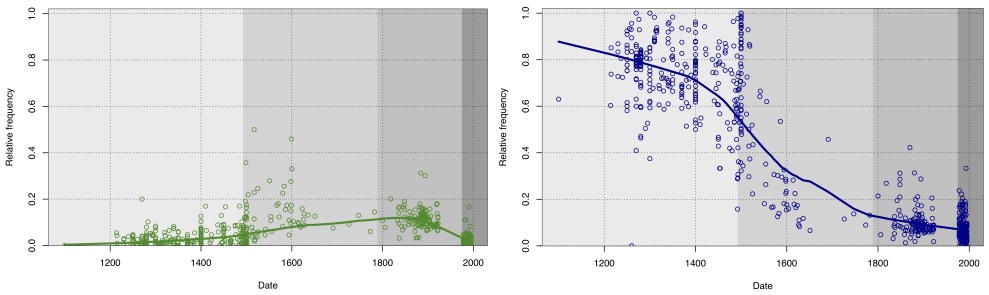

Figure 1: Frequencies of *estar* + participle (left panel) and *ser* + participle (right panel) from the 13th to the 20th century.

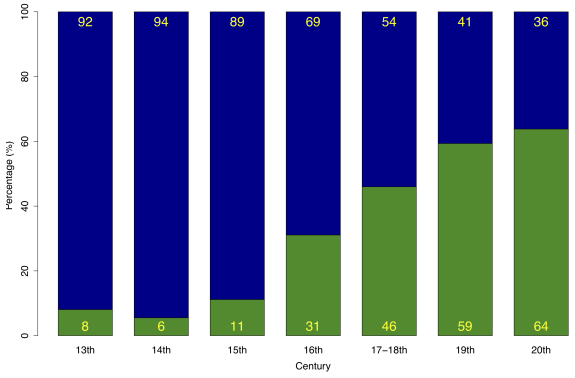

Figure 2: Proportion of *ser* (blue) and *estar* (green) with psychological verbs having an object experiencer from the 13th to the 20th century.

the frequency of use of *estar* with accomplishment and achievement verbs takes place some time after the increase with object experiencer psychological verbs. These results are expected under our hypothesis, as it was presented before.

6 Concluding remarks

With the increasing number of tools and resources to deal with data of historical language varieties, properties of languages in their earliest stages can be better described and explanations about language change can be explored with greater precision. The case study presented in this paper contributes to this research area, in that specific hypotheses to explain lexical extension of a participial construction have been explored in a big corpus, and a deeper understanding of a relatively well-known change has been achieved.

References

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. Linguistic Inquiry Monograph 1, MIT Press.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Batllori, Montserrat and Roca, Francesc (2011). Grammaticalisation of *ser* and *estar* in romance. In D. Jonas (ed.), *Grammatical Change: Origins, Nature, Outcomes*. Oxford University Press.

Beavers, John (under review). Aspectual classes and scales of change. The University of Texas at Austin, Ms.

Bouzet, Jean (1953). Orígenes del empleo de *estar*: ensayo de sintaxis histórica. In *Estudios dedicados a Menéndez Pidal, IV*, pages 56–58. CSIC.

Corfis, Ivy A.; O'Neill, John; Beardsley, Jr. Theodore S. (eds.) (1997). *Early Celestina Electronic Texts and Concordances*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.

Dowty, David R. (1979). *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague PTQ*. Kluwer Academic Publishers.

Gawron, Jean Mark (2005). Generalized paths. In *Proceedings of SALT XV*.

Gawron, Jean Mark (2009). The lexical semantics of extent verbs. San Diego State University, ms.

Harris, Alice C. and Campbell, Lyle (1995). *Historical Syntax in Cross-linguistic perspective*. Cambridge University Press.

Herrera, María Teresa and de Fauve, María Estela González (eds.) (1997). *Concordancias electrónicas del corpus médico español*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.

Joseph, Brian D. (1983). *The Synchrony and Diachrony of the Balkan Infinitive. A Study in Areal, General, and Historical Linguistics*. Cambridge University Press, Cambridge.

Kasten, Llyod; Nitti, John; Jonxis-Henkemens, Wilhemina (eds.) (1997). *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.

Kroch, Anthony (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, **1**, 199–244.

Labov, William (1981). Resolving the neogrammarian controversy. *Language*, **57**, 267–308.

Mendeloff, 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.

Rodríguez-Molina, Javier (2004). Difusión léxica, cambio semántico y gramaticalización: El caso de *haber* + participio en español antiguo. *Revista de Filología Española*, **LXXXIV**, 169–209.

Sánchez-Marco, Cristina; Boleda, Gemma; Padró, Lluís (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, OR, USA. Association for Computational Linguistics.

Sebastián-Gallés, Núria (2000). *LEXESP: léxico informatizado del español*. Edicions Universitat Barcelona.

Sánchez, María Nieves; Herrera, María Teresa; Zabía, María Purificación (2003). *Textos medievales misceláneos*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Sánchez-Marco, Cristina and Marín, Rafael (2011). Generalising paths into psychological verbs. Evidence from Spanish. In *Going Romance XXV*, Utrecht, 7-8 December.

Yllera, Alicia (1980). *Sintaxis histórica del verbo español: Las perífrasis medievales*. Universidad de Zaragoza.

Zwarts, Joost (2006). Event shape: Paths in the semantics of verbs. In *Workshop on the geometric structure of events*, Konstanz, October 7-8, 2004.