Measuring semantic change: The case of Spanish participial constructions

1Cristina Sánchez-Marco, 2Stefan Evert
1Universitat Pompeu Fabra, 2Institute of Cognitive Science, University of Osnabrück
1cristina.sanchez@upf.edu, 2stefan.evert@uos.de

Goal
• Explore different quantitative measures of semantic change
• Contribute to the development of statistical techniques for studies of meaning change

Research questions
Changes in the meaning of Spanish constructions haber ‘have’ + participle, ser ‘be’ + participle, estar ‘be/stay’ + participle, tener ‘have/possess’ + part.

Approach
1. Frequency and productivity
2. Distributional measures of semantic variability
3. Similarity as an explanation for language change

Data
• Corpus of texts from the 12th to the 20th century
• 651 documents, totalling more than 40 million words
• Wide variety of genres and styles
• Enriched with linguistic information (lemma + part of speech)

Results

Frequency

Remarks
• Differences are highly significant (GLM with binomial family and logit link, p < 0.001).
• Also temporal trend within each period.
• Grammaticalization of haber + part and specialization of ser + part.

Open questions
• Controlling effect of genre differences between subcorpora (centuries)?

Density

Remarks
• Shared feature terms (lemmatized)
• Symmetric 7-word window
• Score + sqrt transf. + L2-normalised
• Randomized SVD to 300 dimensions
• Bag-of-word context vectors (Schütze 1998)
• Density = average distance (cosine angle)

Open questions
• Zipf-Mandelbrot LNRE model with “echo” non-randomness adjustment (Baron & Evert 2007).
• Modified to achieve much more satisfactory goodness-of-fit than original “echo” (see table).

Similarity

Remarks
• Same DSM parameters as in density analysis.
• Tener is among the nearest neighbours of haber.

Open questions
• Explain the role that similarity relations play in productivity.
• Neighbours of estar don’t meet expectations.
• Strong influence of DSM parameters. Which model allows linguistically valid conclusions?