

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.sanchezm@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.

Interpretations of the participial constructions in the 13th century

	Perfect	Stative	Verb Passive	Adj Passive	Perfective
<i>haber</i> + participle	+	+			
<i>tener</i> + participle	+	+			
<i>ser</i> + participle	+		+	+	
<i>estar</i> + participle				+	

Interpretations of the participial constructions in the 20th century

	Perfect	Stative	Verb Passive	Adj Passive	Perfective
<i>haber</i> + participle	+				+
<i>tener</i> + participle		+			
<i>ser</i> + participle			+		
<i>estar</i> + participle				+	

1. Is there a significant change in the frequency and usage of participial constructions?
2. If a significant change took place, how and why did it happen?

Approach

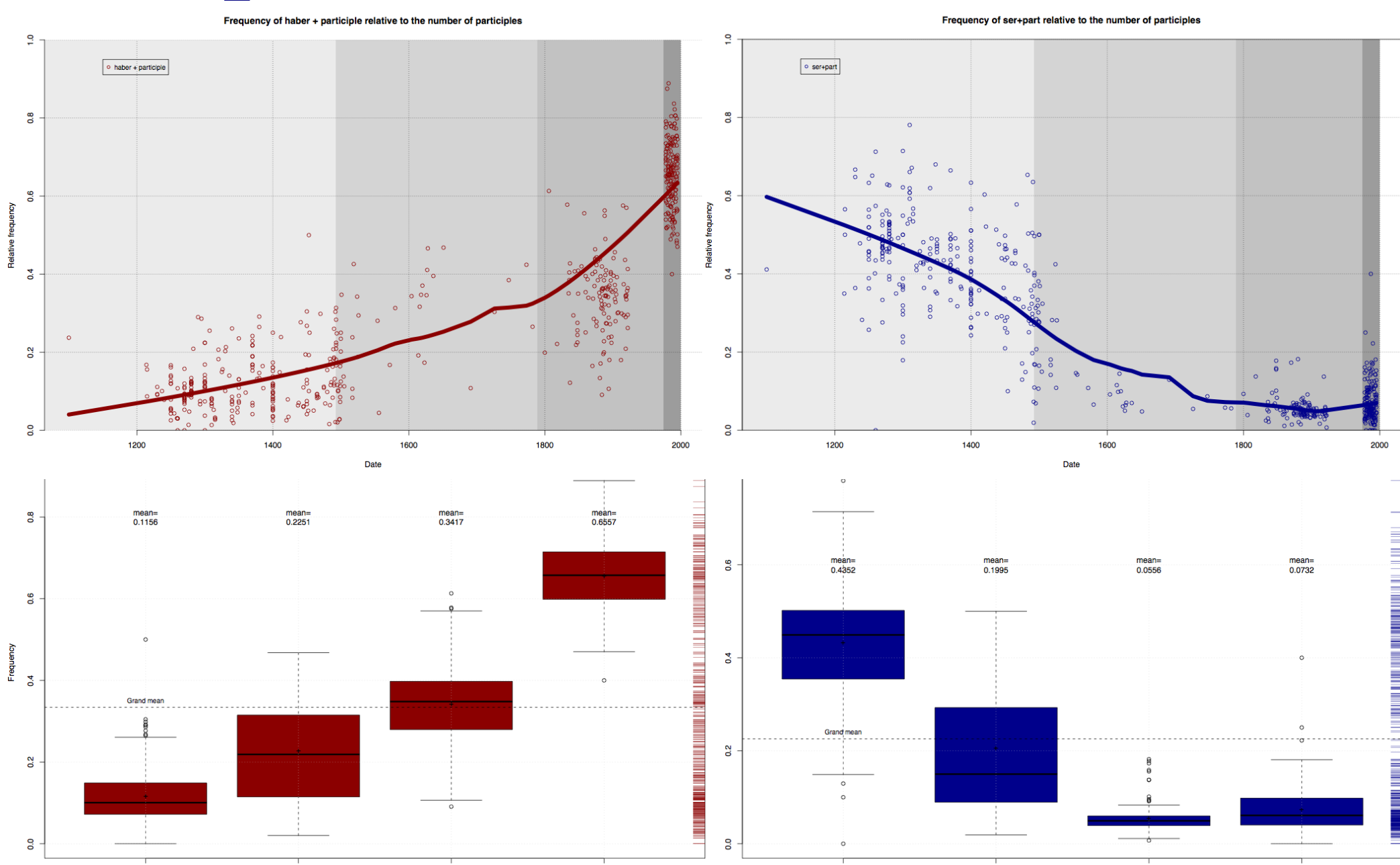
1. Frequency and productivity
2. Distributional measures of semantic variability
3. Similarity as an explanation for language change

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

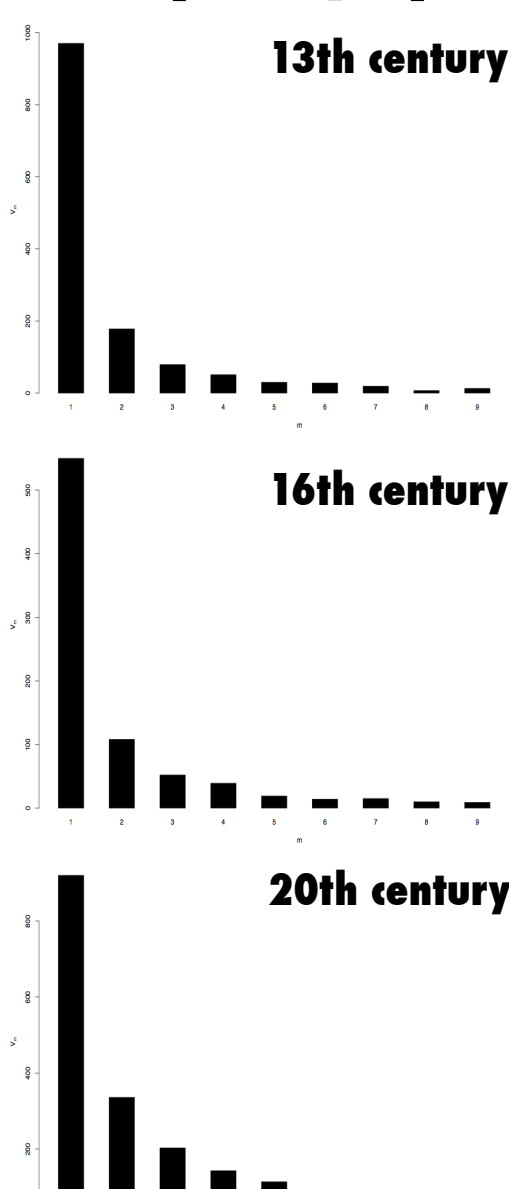
- Differences are highly significant (GLM with binomial family and logit link, $p < .001$).
- Also temporal trend within each period.
- Grammaticalization of *haber* + part and specialization of *ser* + part.

Open questions

- Confounding effect of genre differences between subcorpora (centuries)?

Productivity

Frequency spectrum of *haber*



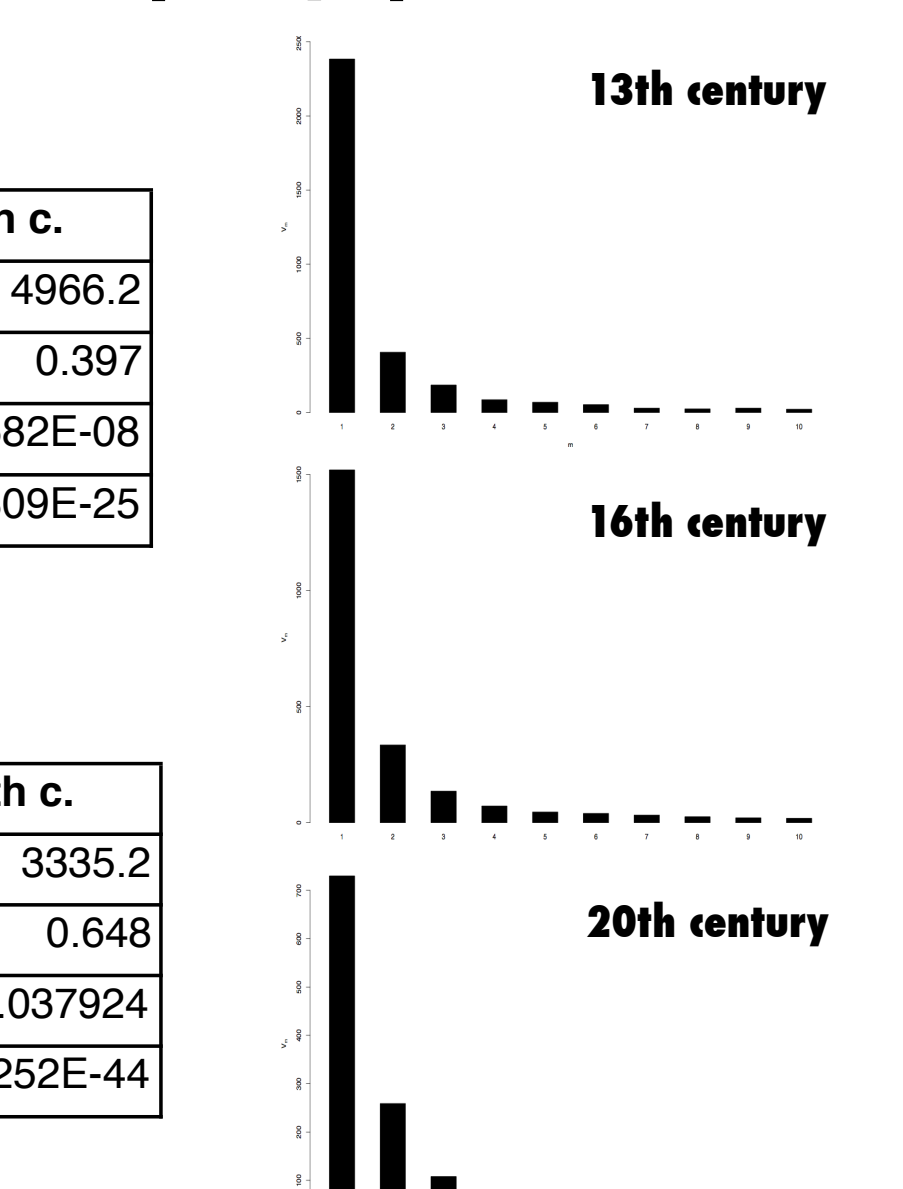
haber + participle

	13th c.	16th c.	20th c.
S	3.423E+12	1.134E+11	4966.2
α	0.615	0.580	0.397
(modified) p	0.240809	0.807386	8.582E-08
(original) p	1.137E-07	3.301E-09	1.809E-25

ser + participle

	13th c.	16th c.	20th c.
S	2.097E+13	24641.0	3335.2
α	0.623	0.683	0.648
(modified) p	2.784E-07	0.773206	0.037924
(original) p	0.022863	3.273E-18	9.252E-44

Frequency spectrum of *ser*



Remarks

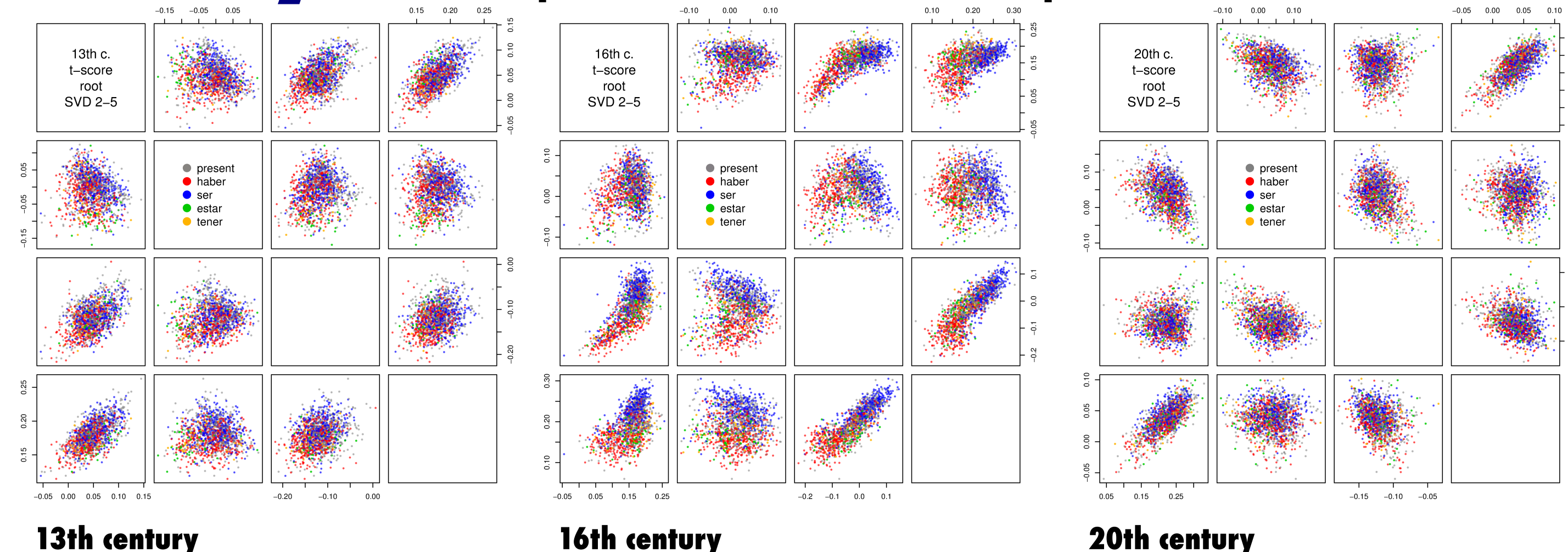
- Zipf-Mandelbrot LNRE model with "echo" non-randomness adjustment (Baroni & Evert 2007).
- Modified to achieve much more satisfactory goodness-of-fit than original "echo" (see table).

Open questions

- Manual correction of data needed (inflated productivity because of lemmatisation errors).
- Estimate sampling error of LNRE parameters.
- Are the observed changes in frequency distributions linguistically meaningful?

Density

Participial constructions in semantic space



	13th c.	14th c.	15th c.	16th c.	19th c.	20th c.
present	21.99°	20.93°	18.69°	20.20°	17.99°	16.62°
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
<i>haber</i>	89.1%	90.8%	90.6%	94.7%	90.1%	93.2%
<i>ser</i>	89.8%	93.7%	96.8%	98.7%	88.7%	91.3%
<i>estar</i>	87.2%	92.3%	95.4%	91.4%	92.1%	96.3%
<i>tener</i>	87.6%	84.3%	90.2%	94.4%	92.9%	97.5%

- shared feature terms (lemmatised)
- symmetric 7-word window
- t-score + sqrt transf. + L_2 -normalised
- randomized SVD to 300 dimensions
- bag-of-word context vectors (Schütze 1998)
- density = average distance (cosine angle)

Remarks

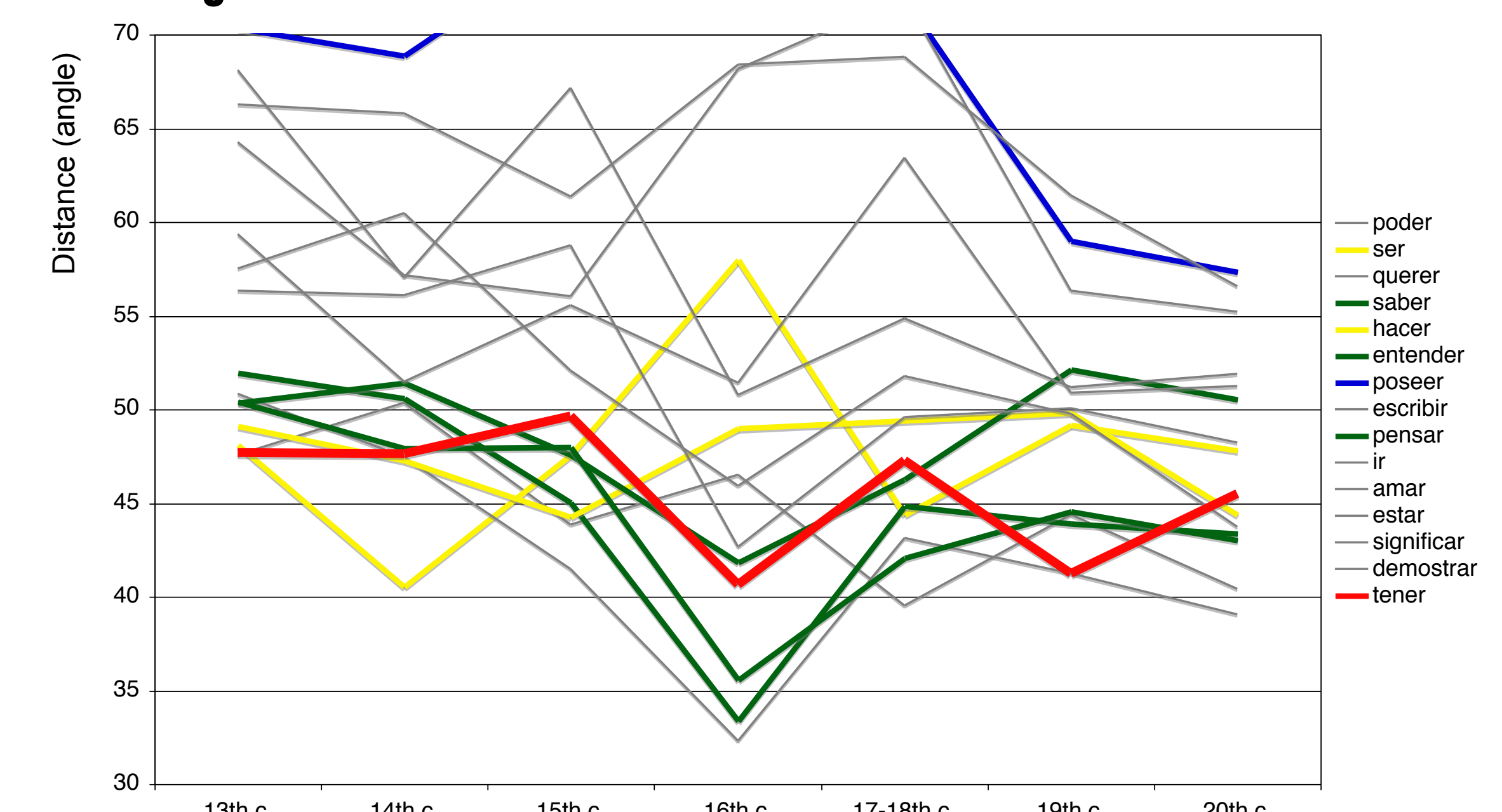
- Verbs in present tense taken as a control category: relative density (present = 100%).
- Increase in the percentage indicates lower semantic density, i.e. less restricted usage.
- Distinct usage patterns in 16th century only.

Open questions

- What is an appropriate measure of density?
- Is it necessary for DSMs to have the same dimensions in order to compare meaning changes in different centuries?
- Does DSM capture the "right" meaning aspects?

Similarity

Neighbours of *haber*



Remarks

- Same DSM parameters as in density analysis.
- *Tener* is among the nearest neighbours of *haber* in the earliest centuries.

Open questions

- Explore the role that similarity relations play in language change.
- Neighbours of *estar* don't meet expectations.
- Strong influence of DSM parameters: Which model allows linguistically valid conclusions?