Tracing semantic change in Portuguese: a distributional approach to *porém*

Patrícia Amaral, Zuoyu Tian & Juan Manuel Escalona Torres
(Indiana University)

In recent years, computational methods based on distributional semantics have been used to detect and explore semantic change in large historical corpora, mostly from English (Hamilton et al. 2016; Tang 2018). This paper uses one such method, word embeddings, to examine semantic shifts in the history of Portuguese, with two goals. First, we seek to contribute to historical studies of Portuguese by using a data-driven methodology that produces replicable and objective measures of change. Second, we show the challenges of using these methods in much smaller corpora than those used in the previous literature, given the limited availability of resources, especially from the medieval period. As a test case, we examine the change in meaning of *porém* 'but, however' (< *por en(de)* 'for this (reason)') from a causal PP to an adversative connective. While this path has been attested cross-linguistically, including in Ibero-Romance (Corominas and Pascual 1980–1991; Mauri and Ramat 2012; Cuenca et al. 2019), to our knowledge this study is the first to rely heavily on quantitative methods to examine this type of meaning change.

Distributional semantics, based on Harris' (1954) seminal work, relies on the assumption that the meaning of a word can be captured by the "company it keeps", i.e. by its distribution in linguistic contexts. In this approach, semantic similarity of lexical items is measured by similarity in contexts of use and semantic change is revealed by a change in the distribution of a word over time, cf. Lenci (2018). This approach provides the field of historical linguistics with tools to move away from approaches mostly based on the linguist's intuition.

However, most studies using distributional methods rely on very large corpora; instead, here we experiment with two historical corpora that are an order of magnitude smaller. We use the CIPM corpus (https://cipm.fcsh.unl.pt//) of Old Portuguese, with ca. 2,4 million tokens, and the COLONIA corpus, which contains texts from the 16th to the 20th century, having ca. 5,1 million tokens (http://corporavm.uni-koeln.de/colonia/). A small corpus such as the CIPM requires careful pre-processing due to orthographic variability, unsystematic punctuation, and a range of editorial annotations. These issues, compounded with the small corpus size, create challenges for word embeddings (e.g. the definition of what counts as a sentence), exacerbating some of the problems previously identified in the literature, e.g. the stability and replicability of the results (Hellrich 2019; Hai et al, under review).

In this particular instance of semantic (and syntactic) change, our aims were (i) to identify a set of similar words revealing the meaning of the word *porém* at different points in time, and (ii) to produce measures of similarity with respect to anaphoric expressions (i.e. its source meaning) and adversative/contrastive markers. Our findings show that the set of most similar words to *porém* in the Medieval Period includes anaphoric adverbs like *então* 'then' and the additive particle *outrossi* 'also, in addition', while the most similar words in the Classical and Modern Periods include *mas, todavia, contudo* 'but, however'. These data confirm the claim that the original semantic meaning of *porém* was not contrastive, but rather additive. Furthermore, similarity measures suggest both the predicted connection to causal markers in Old Portuguese (e.g. *porquanto* 'since' receiving a high score), and a similarity to adverbs with a temporal meaning, both for the older and more recent stages of the language, *entanto* 'during that period of time' and *entretanto* 'in the meantime, meanwhile'. This is worth further exploration, as *entanto* later grammaticalized into an adversative connective (*no entanto*), and *entretanto* today allows for contrastive interpretations, albeit just as pragmatic meanings.

Our paper makes two types of contributions. First, the two measures obtained through word embeddings provide empirical evidence for the predicted semantic shift from anaphoric and causal to contrastive (adversative). Our data also show a semantic connection between *porém* and temporal adverbials that presumably had a similar distribution, hence inviting further investigations on the commonality of contexts between causal and temporal adverbials. Second, from a methodological point of view, our work shows that applying word embeddings to such small data sets requires a careful balance involving frequency, accuracy, and corpus size. The higher the minimum value for the word frequency, the more accurate the word embedding model is. However, given the small size of our corpus, raising the frequency cut-off could mean a significant loss of data.

References:

Corominas, J. and Pascual, J. A. (1980-1991) *Diccionario crítico etimológico castellano e hispánico*, 6 vol., Madrid, Gredos.

Cuenca, M. J., Postolea, S. and Visconti, J. (2019) Contrastive markers in contrast. *Discours* (online) 25.

Cunha, A. G. da. (2014) *Vocabulário histórico-cronológico do português medieval*, Rio de Janeiro, Fundação Casa de Rui Barbosa.

Hai, H., Amaral, P., and Kübler, S. (under review) Word embeddings and semantic shifts in historical Spanish: Methodological considerations.

Hamilton, W.L., Leskovec, J., and Jurafsky, D. (2016) Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1489-1501, Berlin.

Harris, Z. (1954) Distributional structure. *Word,* 10(2-3): 146-162.

Hellrich, J. (2019) *Word Embeddings: Reliability and Semantic Change*. PhD dissertation, Jena University Language and Information Engineering Lab.

Huber, J. (1986 [1933]) *Gramática do Português Antigo*. Lisboa: Fund. Calouste Gulbenkian.

Lenci, A. (2018) Distributional models of word meaning. *Annual Review of Linguistics*, 4: 151-171.

Mauri, C. and A. G. Ramat (2012) The development of adversative connectives in Italian: Stages and factors at play. *Linguistics* 50(2): 191-239.

Silva, R. V. M. e. (1994) *O Português Arcaico. Morfologia e sintaxe*. São Paulo: Contexto.

Tang, X. (2018) A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5): 649-676.