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

Patrícia Amaral, Zuoyu Tian & Juan M. Escalona Torres

Abstract

This study takes a computational approach to language change by employing the method of word embeddings. Our goal is two-fold. First, we seek to contribute to the study of historical linguistics using data-driven methodology that produces replicable and objective measures of semantic change. Second, while previous studies have used large sized corpora when using word embeddings, we present the challenges that come with analyzing a smaller text sample from Medieval Portuguese. 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.

Semantic and Syntactic Change

Porém. There is syntactic change from a PP headed by *por* to a function word and semantic change from causal to adversative meaning, attested cross-linguistically (Mauri and Ramat 2012; Cuenca et al. 2019).

Medieval Portuguese (CIPM)

(1) …mas por que os seus feytos nõ foron muyto assiinados pera contar em esta estoria **poren** tornaremos a contar de hercolles que foy o homem que mais feytos assiinados fez (*Crónica Geral de Espanha de 1344*)
'but since their feats were not remarkable (enough) to be told in this narration, **for this reason** we will narrate those of Hercules, who was the man who did the most remarkable feats'

(2) …todos outorgavam de seer em tall feito, mas nuhuum nom sse atrevia de seer o primeiro; e o Comde bem entemdia que de taaes pessoas nom era mui seguro, nom damdo **porem** a entender nada; mas seu gramde estado e aguardamento de muitos…o fazia segurar de todos (*Crónica de D. João I, Fernão Lopes,* 15th c.)
'all declared to be willing to do that [kill the Count], but none dared to be the first; and the Count knew well that he could not trust those people, **for this reason/but** he didn't reveal anything, but his large guard protected them from all'
[Ambiguous example]

Classic-Modern Portuguese (COLONIA)

(3) Disse-lhe o criado o que passava, quietou-se algum tanto, **porém** não deixou de ficar queixoso e dando suspiros. (*A vida de Frei Bartolomeu dos Mártires*, 1606)
'The servant told him what was happening, [and] he got a bit calmer, **but** he was still complaining and sighing'

Hypothesis: the change took place in negative contexts (e.g. Said Ali 1971; Mauri and Ramat 2012), cf. (2).

Distributional semantics. Distributional semantics assumes the distributional hypothesis: words with similar meaning occur in similar contexts (Harris 1954; Lenci 2018). 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. The starting point to represent word meaning in context is to take co-occurrence counts of words in a text. The meaning of a word consists of a vector of numbers corresponding to frequency of co-occurrence, e.g. *espada*: <…,30, 10,…,1, o,…>. Since these tables are sparse matrices with many dimensions, there are dimensionality reduction techniques. The most recent models use neural networks.

	…	baínha	mão	…	denis	vila	…
espada		30	10		1	0	
braço		10	52		0	0	
rei		4	5		30	10	
casa		0	0		1	35	

Research Goals and Methods

Research goals. In examining the semantic change undergone by *porém* we seek to do the following:

1. Identify a set of similar words revealing the meaning of *porém* at different points in time (providing **evidence for semantic and syntactic change**);
2. Produce **measures of similarity** with respect to the source meaning and the new meaning;
3. Develop appropriate **methodological tools** in order to address the challenges of **small corpora of historical data** for a distributional analysis with word embeddings.

Research hypotheses. We predict that the semantic neighbors of *porém* in Medieval Portuguese will be causal and anaphoric expressions. On the other hand, as *porém* acquires adversative functions in Classical and Modern Portuguese, neighbors in Colonia will increasingly be adversative expressions.

Data. The data used in this study comes from two historical corpora (see Table 1).

Table 1. Data

Corpus Informatizado do Português Medieval (CIPM)	*Colonia: Corpus of Historical Portuguese*
2,888 documents	100 documents
2.8+ million words	5.1+ million words
13th to 16th century	16th to 20th century
https://cipm.fcsh.unl.pt//	http://corporavm.uni-koeln.de/colonia/index.html

Preprocessing. The texts from the corpora are digitized versions of paleographic editions of Medieval and Classical Portuguese. As such, these text contain a great deal of metalanguage annotations that interfere with language processing. We removed all annotations and punctuation.

Normalizing. The texts in CIPM exhibit spelling variation. Words with different variants are detected by the software as different words (e.g., don, dom, dõ). To reduce such variation, we normalized consistent patterns (e.g., daquela/d'aquela/d' aquela/daquella > de aquela).

Word Embeddings

What is a word embedding? A word embedding is a computational method that objectively measures a word's neighboring context within a large data sample. From a historical standpoint, a change in a word's neighboring context may reflect semantic and syntactic changes over time. The word embedding analysis allows the researcher to examine language change by comparing a word's recurring contexts during different time periods. In this study, we use Skip-Gram Negative Sampling (SGNS) to build the word embeddings.

What does a word embedding look like? First, we plot the target words. In this example, we use the set of common nouns *rey, braço, espada, casa*. Closeness in space means similarity in meaning.

Figure 1. Similarity between the words in CIPM

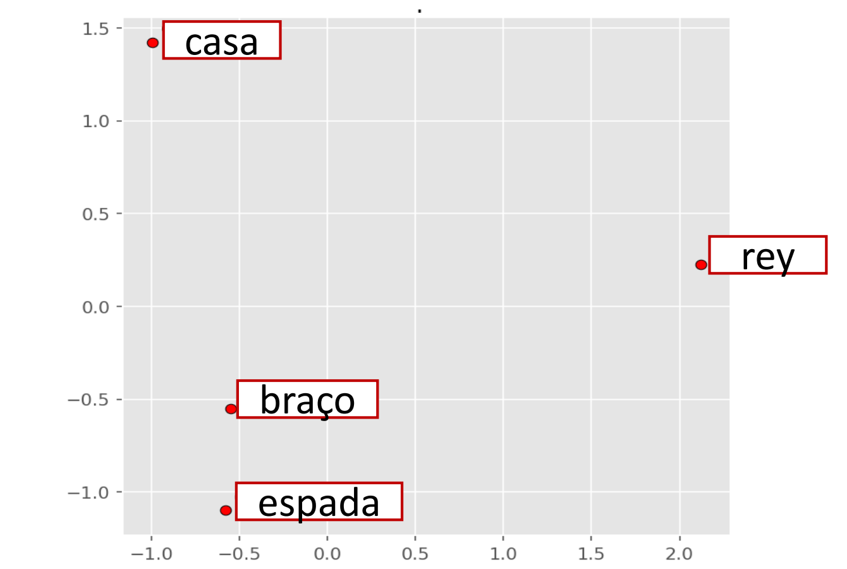
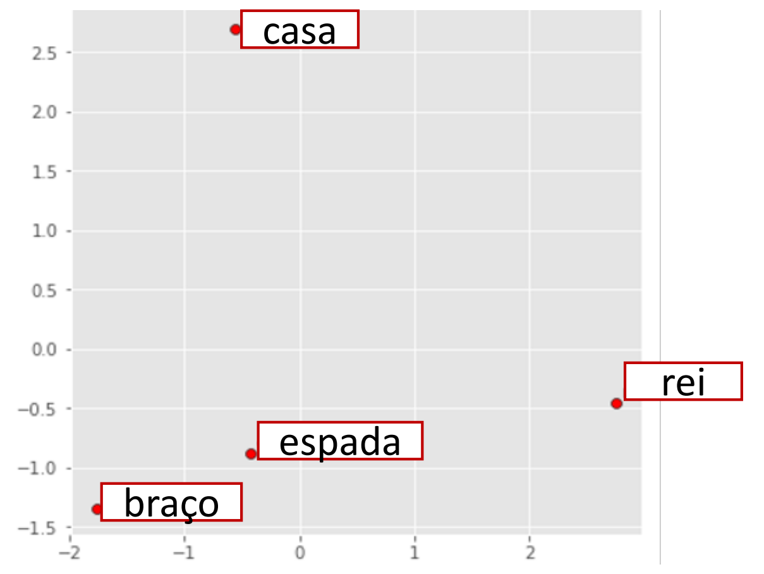


Figure 2. Similarity between the words in Colonia



Second, we plot the clusters of the 30 most similar words for each target word. In this example, we show the clusters for Set 2.

Figure 3. Neighbors of *braço, espada, rei* in CIPM

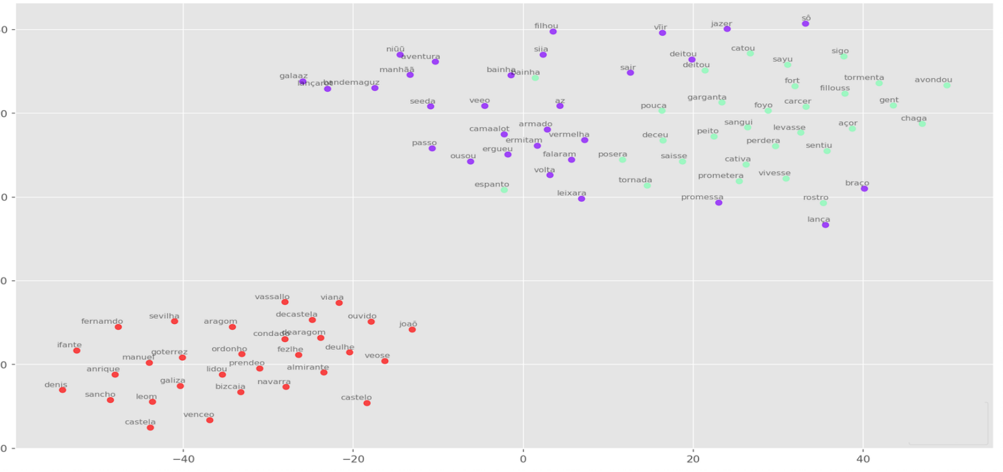
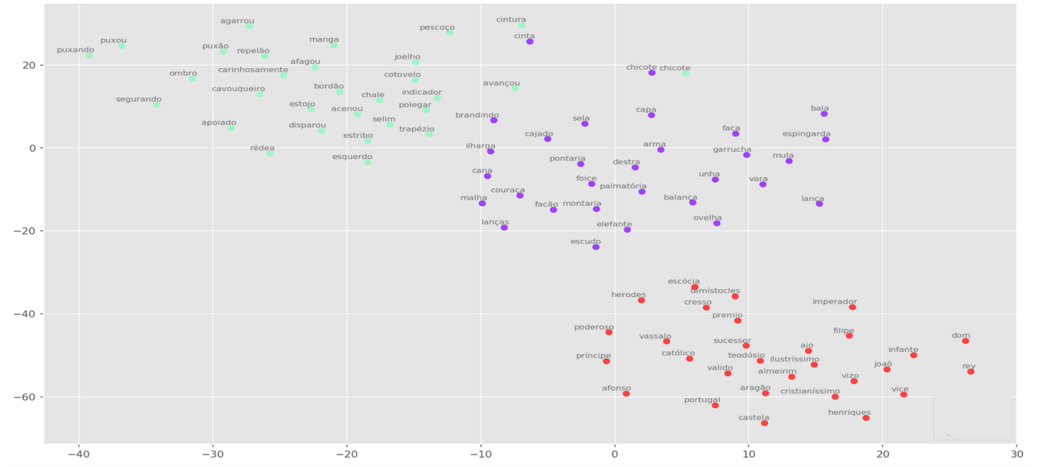


Figure 4. Neighbors of *braço, espada, rei* in Colonia



Assessment of word embedding models

To evaluate the performance of different embedding settings, we created an **analogy test dataset for historical Portuguese**. Contemporary analogy tests are not appropriate due to lack of shared vocabulary. The test includes: N Sing-Pl, Antonyms, N Gender, V Infinitive-Gerund, Concept categorization, Outliers, etc. Our preliminary results show that accuracy in the analogy test tends to coincide with judgements about models based on previous research.

We also observe the reduction in the data because of frequency cut-offs. For CIPM, about 85% tokens of the whole corpus are utilized, 8565 unique words are retained in the vocabulary (4% of original 193196, drops 184631). For Colonia, about 92% tokens of the whole corpus are utilized, 15936 unique words are retained in the vocabulary (11% of original 133070, drops 117134).

Selected References

Cuenca, M. J., Postolea, S. and Visconti, J. (2019). Contrastive markers in contrast. *Discours* (online) 25.
Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Association for Computational Linguistics* (ACL), 1489–1501. Berlin, Germany.
Harris, Z. (1954) Distributional structure. *Word*, 10(2-3): 146-162.
Hu, H., P. Amaral and S. Kübler (under review) Word embeddings and semantic shifts in historical Spanish: Methodological considerations.
Lenci, A. (2018) Distributional models of word meaning. *Annual Review of Linguistics*, 4: 151-171.
Mauri, C. and Ramat, A. G. (2012) The development of adversative connectives in Italian: Stages and factors at play. *Linguistics* 50(2): 191-239.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ.
Said Ali, M. (1971). *Gramática histórica da língua portuguesa*. Rio de Janeiro: Livraria Acadêmica.
Silva, R. V. (1994). *O Português Arcaico: Morfologia e sintaxe*. São Paulo, Brasil: Contexto.

Results

In order to test our hypotheses, we chose 4 words in each category, causal, anaphoric, adversative, for both periods (based on Huber 1933 and Silva 1994).

Figure 5. Similarity scores of *porém* per word category for both time periods. The similarity score for each word is the average of 5 models. The results of a t-test indicate statistically significant differences in semantic similarity by word group.

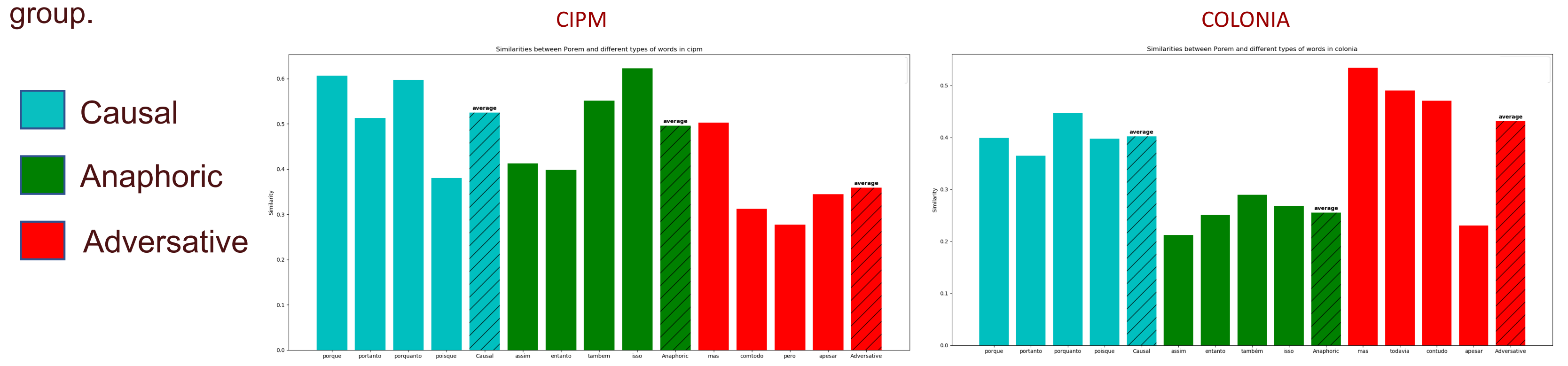


Table 2. T-test results for similarity scores in CIPM and COLONIA by word groups

Dataset	Group	*p* value	Significance
CIPM	Causal vs Anaphoric	0.46772	
CIPM	Causal vs Adversative	0	**
CIPM	Anaphoric vs Adversative	0.00121	**
Colonia	Causal vs Anaphoric	0	**
Colonia	Causal vs Adversative	0.31328	
Colonia	Anaphoric vs Adversative	0.00002	**

Table 3. Ten similar words to *porém* in CIPM (after selection, average similarity scores from 24 models)

Word	Similarity Score (average)
comtrairo	0.727738
portamto	0.722877
geeralmente	0.719007
sollamente	0.718629
embargamdo	0.710035
neçessario	0.709037
dello	0.696267
postoque	0.695050
senom	0.682459
isso	0.656867

Table 4. Ten most similar words to *porém* in COLONIA

Word	Similarity score
entretanto	0.557456
mas	0.545567
todavia	0.478458
portanto	0.462748
obstante	0.461777
contudo	0.456721
porquanto	0.451362
verossímil	0.438663
porque	0.436483
agravar	0.432998

Parameters of the model*: min_count:20; size: 200; epoch: 5; window size: 7; skip gram model*

Table 5. Examples of contexts containing neighbors of *porém* in CIPM and COLONIA

(1) e **nõ** tan sollamente esto mas que se partiria logo da terra (*Crónica Geral de Espanha*, 14th c.)

(2) cuidadas bem taaes rrazõoes **nom** embargamdo seu ardido coraçom e boa voomtade (*Crónica de D. João I*, 15th c.)

(3) que som navios que **nom** podem levar senom pouca gemte homde compria de levar muita (*Crónica de D. João I*, 15th c.)

(4) **mas** esto era muito pello comtrairo **ca** as gemtes da çidade todas eram dhuũ acordo (*Crónica de D. João I*, 15th c.)

(5) Esta última idéia lhe sorria mais; entretanto não tomou nenhuma resolução definitiva (*O Guarani*, 19th c.)

Conclusion

Research Goal #1. The change in the semantic and syntactic domains of similar words provides **evidence for syntactic and semantic change**. In COLONIA, the neighbors of *porém* are adversative and some causal expressions. The word embeddings in Medieval Portuguese show causal and anaphoric expressions but also expressions introducing contrast and occurring in negative contexts. This provides support for the hypothesis about the role of negation in the development of adversatives.

Research Goal #2. In CIPM, *porém* is **significantly more similar to causal and anaphoric expressions** than to adversatives. In COLONIA, the similarity regarding anaphoric expressions is lower than with causal and adversative expressions; **its anaphoric function has significantly decreased**. There is no significant difference between the similarity to causal and adversative connectives.

Research Goal #3. For a small corpus like the CIPM, working with the **intersection of neighbors from multiple models or a ranking algorithm** (Hu et al. under review) yields better results. A selection process is required due to the instability of the output (a property of the models which is exacerbated with small data sets).

Conclusion. These methods are reliable indicators of semantic similarity that do not depend on the linguist's intuition (or at least not exclusively, as most traditional accounts to semantic change). Since word embeddings use the entire corpus, we get insight into broader semantic relations in the language at different time points.

FUNDED BY: NEW FRONTIERS IN THE ARTS AND HUMANITIES AND IDAH