WORDS THAT SHIFT ARE LEARNED LATER: A STUDY ON THE RELATION BETWEEN DIACHRONIC SEMANTIC CHANGE AND AGE OF ACQUISITION

Giovanni Cassani (Tilburg University), Federico Bianchi (Bocconi University) & Marco Marelli (University of Milan – Bicocca)
g.cassani@tilburguniversity.edu

We study the relation between Age of Acquisition (AoA) and semantic change, quantifying it using word embeddings extracted from the Corpus of Historic American (CoHA) English (Davis, 2010), a diachronic corpus of American English spanning the period 1800-2009. Previous research (Monagahn, 2014; Monaghan & Roberts, 2019) looked at the relation between language evolution and AoA, but quantified language evolution through cognate proliferation and rate of lexical substitution for a limited set of words. We extend this line of work by exploring semantic change, using a highly scalable corpus-based method to quantify it for thousands of words. Our results show that the relation between AoA and language change also holds for semantic shift on a smaller time scale than previously established.

We first bin the CoHA in six slices containing a similar number of tokens and then align embedding spaces across time slices using Temporal Word Embeddings with a Compass (TWEC, Di Carlo et al. (2019)). TWEC extends the Continuous Bag of Words model (CBoW, Mikolov et al. (2013)) and aligns different vector spaces to a shared coordinate system in such a way that the different embedded slices can be compared.

We consider two different measures of semantic change, tracking the coherence of the embeddings of a word over time (vector coherence, VC) and the local neighborhood coherence (LNC) of a word over time, in line with the work by Hamilton et al. (2016). Words with high VC have similar embeddings over time while words with high LNC have a consistent relation with their semantic neighbors across time slices. For each measure, we also compute a quasi-random baseline to ensure that any observed effects are robust.

We fit regression models using AoA as the dependent variable and focus on the 8296 words for which we have AoA norms (Kuperman et al., 2012), concreteness ratings (Brysbaert et al., 2014), frequency estimates from SUBTLEX US (Brysbaert and New, 2009) and which occurred in all temporal slices with a frequency count >= 25 in order to estimate reliable embeddings. We add our target measures separately to a baseline model including log frequency, OLD20 (Yarkoni et al., 2008), length in letters, concreteness, dominant Part of Speech, and frequency in the CoHA. We then compare the degree to which they improve the fit over the baseline model, observing that both explain additional variance over the base model and that words with lower rates of semantic change are acquired earlier. In spite of the fact that VC has a stronger correlation with AoA, LNC explains a higher proportion of extra variance over the base model. Random baseline measures, on the contrary, have far weaker correlations with AoA.

Qualitative analyses show that words with high VC tend to be basic level, concrete concepts (e.g. *dress*, *water*, *fish*, *door*). On the contrary, among concepts with high LNC we primarily see words referring to body parts, parts of the day, time expressions, family relations, trees and fruit. Phonosymbolic words and words including phonaestemes also tend to have high LNC (e.g. *howl*, *shriek*, *glimmer*, *fling*, *snore*, …), suggesting that words with high LNC may capture statistical regularities in form-meaning relations. Future studies should investigate this possibility further.

Summing up, we report evidence of a robust relation between semantic change (quantified with a scalable, corpus-based method for thousands of words) and AoA, with early acquired words showing higher semantic coherence over time, beyond the effects of standard covariates. We also show that different measures of semantic change highlight different aspects of this phenomenon, characterizing it as a multidimensional construct. On the basis of preliminary qualitative analyses, we hypothesize that this relation signals that words with consistent meanings over time refer to aspects of the world which are more salient to children and serve to structure later vocabulary learning, as they denote stable concepts, whose meaning is less culturally determined.

Technical details of the computational model:

The TWEC (https://github.com/valedica/twec) model is implemented using the *gensim* (https://radimrehurek.com/gensim/) library and is based on the CBoW model, which is instantiated in a neural network with one hidden layer. The CBoW model uses two matrices to learn word embeddings, a target matrix and a context matrix. TWEC exploits this aspect by first training a general embedding using the whole corpus, ignoring the time dimension. This embedding space is the *compass*, i.e. a general representation to which the embedding spaces derived from each time slice can be aligned. The context matrix of the compass is extracted and used to initialize (and freeze) the second matrix of a slice-specific CBoW embedding space. This approach ensures that all slice-specific embedding spaces share the same context matrix, making the slice-specific embeddings aligned.

The reported analysis is performed on word embeddings with 100 dimensions, learned using a window of 5 words to the left and to the right of the target word. However, we performed a small grid search exploring windows of 3 and 20 words, as well as embeddings with 40 dimensions. Results do not depend on these hyper-parameter choices. Parameter optimization is carried out with a learning rate of 0.025 and 10 negative samples. We use 5 iterations to train the compass embeddings and 5 iterations to train the slice specific embeddings. We initialize all the other hyper-parameters using the default settings provided by the library.

References:

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019, July). Training temporal word embeddings with a compass. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 6326-6334).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 2116).

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, *44*(4), 978-990.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. *Cognition*, *133*(3), 530-534.

Monaghan, P. and Roberts, S. G. (2019). Cognitive influences in language evolution: Psycholinguistic predictors of loan words borrowing. Cognition, 186, 147-158.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic bulletin & review*, *15*(5), 971-979.