Distributional Semantic Models for Vocabulary Acquisition

Raquel G. Alhama[1], Caroline Rowland[1,4] and Evan Kidd[1,2,3,4]
[1]Max Planck Institute for Psycholinguistics; [2]The Australian National University; [3]ARC Centre of Excellence for the Dynamics of Language; [4]Donders Institute for Brain, Cognition & Behaviour
rgalhama@mpi.nl

Building a lexicon is a fundamental component of language acquisition. To navigate the immense hypothesis space of word-meaning mappings, children appear to draw on specific preferences for certain mappings (e.g., via biases like mutual exclusivity), in addition to tracking the co-occurrence of words and referents across multiple situations. In our work, we study another source of information for disambiguating word-meaning mappings: the linguistic context around a word. The context of a word provides information about the meaning of a word; for instance, we may infer that 'niwe' is a beverage if it appears often in the context of 'glass' and 'drink'. This idea is the cornerstone of Distributional Semantic Models used in Computational Linguistics; here we apply these algorithms to language acquisition.

We investigate how distributional learning bootstraps the acquisition of words with two modelling approaches: context-counting and context-predicting models. These two approaches implement different algorithms for deriving word representations, based on either *counting* co-occurrences between words and contexts, or *predicting* which words are likely to appear in the context of a given word. The resulting word representations preserve semantic similarity: two representations are more similar if they are semantically related (see Appendix 2 for more details).

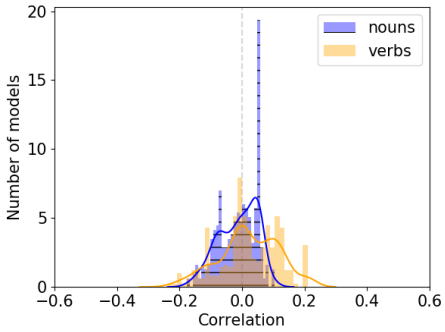
We trained a range of models using both approaches, systematically varying their parameter configurations (e.g. the number of words that are considered as context to a given word). Notably, the models were trained on child-directed speech from CHILDES (English; 0-60 months). We then counted the number of semantically close neighbours of each word, based on the similarity of the representations, and evaluated the models on how well they predicted age of acquisition norms (AoA) by correlating the number of neighbours against AoA norms from Wordbank (Frank et al., *J. of Child Language*, 2017).

Our simulations allowed us to investigate how specific distributional cues are reflected in the emerging word representations. We find that, while the correlations between AoA norms and semantic density in context-counting models cluster around zero, context-predicting models yield much stronger, mostly positive significant correlations (see Figure 1). This shows that context-counting and context-predicting models build semantic spaces that differ in this evaluation metric, yielding opposite predictions. The better fit of the context-predicting model points to a predictive process that children may employ to progressively refine the representations of words until they are successfully acquired.

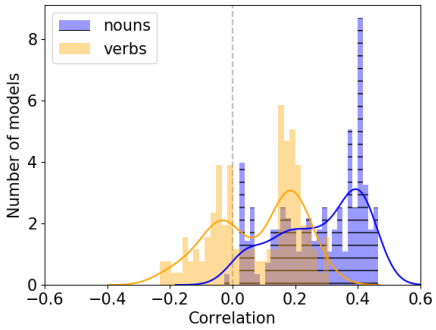
The simulations with the context-predictive models reveal several further insights into vocabulary acquisition. Firstly, words in denser semantic neighbourhoods are associated with delayed acquisition, supporting the intuition that acquiring fine-grained meaning requires more data. Secondly, a word's local context is most informative: a window size of 1 provides the best fit to the data, which aligns with the idea that children's constrained memories are advantageous in acquisition (Johnson&Newport, *Cogn. Psychology*, 1989). Finally, we find that raw frequency plays an important role: frequent words tend to have fewer neighbours, and are acquired earlier. This suggests that greater exposure makes words have a more distinct representation from that of their neighbours, which would act as competitors.

Overall, our research reveals the prerequisites for a distributional model of language acquisition: word learning is supported by a mechanism that operates over a very local context and progressively refines word representations until their meaning is clear enough for the word to be incorporated in the emerging vocabulary.

Appendix 1: Figures



(a) Context-counting.



(b) Context-predicting.

Figure 1: Frequency distributions of Pearson's *r* correlations between semantic density and AoA, for a range of model configurations of each approach.

Appendix 2: Details of Computational Models

Distributional Semantic Models provide word representations derived from linguistic productions. These representations reflect information of the context around each word, limited to a certain window size. The word representations are in fact numerical vectors: in their simplest form, these vectors encode the number of co-occurrences between the represented word and all other words in the corpora. However, in most cases these vectors are either later refined with mathematical transformations (as in context-counting models), or they are directly *learnt* from the data with neural network models (as in context-predicting models).

What makes these representations interesting is that they encode a notion of semantic similarity: the vectors can be seen as points in a multidimensional space, where words that are close to each other in this space tend to be semantically related. For instance, the vector for the word 'fork' is likely to be close to the vector for 'spoon', since both words appear in similar contexts (and hence will have similar representations).

We implement our context-counting model by first gathering co-occurrence counts from the corpora and then transforming these counts with Positive Pointwise Mutual Information (PPMI), a measure known to help avoid the skew of raw co-occurrences. This results in a high-dimensional matrix of PPMI counts, which we then compress using Singular Value Decomposition (SVD), a matrix factorization procedure that reduces the number of dimensions by removing redundancy.

In Computational Linguistics, context-counting models have largely been replaced by better-performing context-predicting models. These models are trained in the task of predicting words that may appear in the context of a given word. The internal weights of the network are taken as the vector word representations because they preserve semantic relatedness (and hence words with similar representations are likely to be related). As a context-predictive model, we use the *skipgram* version of the *word2vec* framework (Mikolov et al., *CoRR*, 2013; *NeurIPS*, 2013), which is a feed-forward neural network model trained to predict, for a given word, the probability of other words appearing in its context.

Given the greater degrees of freedom of neural networks, the latter approach may exploit the co-occurrence frequencies in a more flexible way, depending on how these aid the prediction of which words will appear in the context of a given word. Thus, these two approaches implement different processes for using word context (even though specific parameter settings can result in both models performing equivalent computations (Levy et al., *TACL*, 2015), our results show a tendency to provide different representations, in terms of semantic density).