# Exploring potential gender stereotypes in the distributional semantics of child-directed speech

**Benjamin E. deMayo (bdemayo@princeton.edu)**
Princeton University

## Abstract

Abstract: In three analyses, I explore whether gender stereotypes might be present in the distributional semantics of the CHILDES corpus (MacWhinney 2000), a large compendium of transcribed conversations between caregivers and their children, by training 2 commonly-used word embedding models on the corpus. In the first analysis, I examine the correlation between the gender valence of individual word vector representations in the two word-embedding models. In the second analysis, I relate gender valence in the word vector representations of individual words to human ratings of those words' gender valence. In the third analysis, I examine whether specific stereotypical associations with gender are detectable in the vector space representation of the words.

**Keywords:** gender bias; word-embedding models; distributional semantics.

## Introduction

Gender is a highly salient social category that develops within the first few years of life and maintains its importance across the lifespan (Ruble, Martin, & Berenbaum, 2006). Gender stereotypes, which can be thought of as characteristics that are believed to be true of a gender category as a whole, have their origins in toddlerhood, but become more rigid in the preschool years into middle childhood (Halim & Ruble, 2010). How children form concepts of gender, as well as the stereotypes that are linked to those concepts, has long been a subject of research, with some researchers proposing that language input to children could have a consequential impact on children's conceptualizations of gender categories.

Broadly speaking, two theoretical approaches have attempted to explain how language input to children might shape their gender concepts. One approach has emphasized the communication of knowledge from adults to children in a "top-down" fashion, in which children hear statements that explicitly communicate information about groups, such as generic statements (e.g., "girls are good at reading"; Gelman, Ware, & Kleinberg, 2010). Another approach, which I focus on here, emphasizes how children could pick up on subtle cues about gender concepts and stereotypes from the statistics of their language input in a "bottom-up" fashion. In other words, children could learn that words corresponding to particular activities, traits, occupations, and other characteristics are themselves gendered by virtue of the other words with which they co-occur. This latter approach therefore shares an intimate link with the computational linguistic subfield of distributional semantics, which seeks to characterize how the meaning of linguistic items is related to how those items are distributed in large bodies of text.

Several studies have leveraged the tools of distributional semantics to examine whether gender stereotypes are appreciable natural language corpora; some of these studies focus specifically on language that would likely be heard by children. The general strategy used by these studies has involved taking large bodies of text (usually those with several million tokens, though this has not always been the case Lewis, Borkenhagen, Converse, Lupyan, & Seidenberg, 2020) and using them to train word embedding models, which generate representations of individual word types in a high-dimensional vector space based on each word type's co-occurrence with other types. The key assumption in such a strategy is that words that frequently co-occur will have similar meanings. Once vector representations of words are obtained, cosine distances between individual lexical items in the vector space are calculated as a proxy of semantic similarity, allowing researchers to examine whether words' vector representations show patterns of similarity to other words that might be expected given prevalent societal stereotypes (e.g., that the word "doll" is closer to the word "girl" than it is to the word "boy"). This general analytic framework has been used to argue that gender stereotypes are present in the distributional structure of large bodies of naturalistic text, including web-based corpora, children's books, and transcripts of films and television shows (Bhatia & Bhatia, 2021; Caliskan, Bryson, & Narayanan, 2017; Charlesworth, Yang, Mann, Kurdi, & Banaji, 2021; Lewis & Lupyan, 2020).

In this work, I extend prior findings by examining the human-like gender-stereotypical biases that might emerge in the vector representations obtained from training word embedding models on a body of child-directed speech. Specifically, I use transcripts between caregivers and children between the ages of 1 and 3 years old from the North American English corpora in the Child Language Data Exchange System [CHILDES; MacWhinney (2009)] and extract vector-space semantic representations using 2 commonly-used word embedding models, word2vec (Mikolov, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014).

# Method

## Data preprocessing

Child-directed language was sourced from all of the North American English transcripts in the CHILDES corpus (MacWhinney, 2009). Transcripts were obtained using the `childes-db` API, which allows researchers to access transcript utterances in a tabular format that includes metadata about each utterance, including its speaker's role (parent, grandparent, child, etc.), the gender of the child in the conversation, and the lemmatized "stem" of the utterance (Sanchez et al., 2019). From this tabular data, the stems of utterances from mothers, fathers, grandparents and adults were extracted and concatenated to create the training data, which contained X conversations from Y dyads including Z children, and was comprised of A word tokens and B word types.

## Word embedding model training

Two common word embedding models were used to obtain vector-space representations for words in CHILDES. The first was Word2Vec (Mikolov et al., 2013), which uses a 2-layer neural network to predict a given word in a sentence given its surrounding words (continuous bag of words approach, CBOW) or vice versa (skip-gram approach) and derives vector-space representations of each word based on the neural network weights between the input layer and the single hidden layer of the network. The second was GloVe (Pennington et al., 2014), an unsupervised learning algorithm which takes as input a sparse matrix encoding the co-occurrence frequency of each pair of lexical items in a corpus, and which learns vector representations for these items, such that the inner product between two vectors closely approximates a logarithmic transformation of the probability that those two lexical items co-occur in the text. For our purposes here, the two techniques have the same goal of extracting vector representations of words that are semantically meaningful, even though GloVe's learning strategy emphasizes co-occurrence probability between pairs of words more than Word2Vec, which centers more on the semantic contexts that words appear in. In the following analyses, the context window of each word embedding model is set to 5 words in both directions from a target word. Word representations derived from GloVe are vectors in a X-dimensional space and those from Word2Vec are in a Y-dimensional space.

# Analyses

## Analysis 1: Broad comparison between Word2Vec and GloVe

The first, and most broad, analysis is a coarse indication of whether Word2Vec and GloVe are capturing roughly similar semantic information for words in the CHILDES corpus, particularly as it concerns individual words' gender valence. In this analysis examine whether the gender valence of a word's representation in Word2Vec is associated with the gender valence of the same word's GloVe vector representation. The degree to which a word is gendered is operationalized here as the average cosine distance between the word's vector representation and the vector representations of each of a set of "anchor words" which correspond to the concept of "boy" or "girl." For this analysis, anchor words were borrowed from Lewis et al. (2020) and were as follows: girl words = {girl, woman, sister, she, her, daughter}; boy words = {boy, man, brother, he, him, son}. For each word vector obtained from both GloVe and Word2Vec, we can calculate this average cosine distance to the set of anchor words representing the broader concepts of "boy" and "girl."

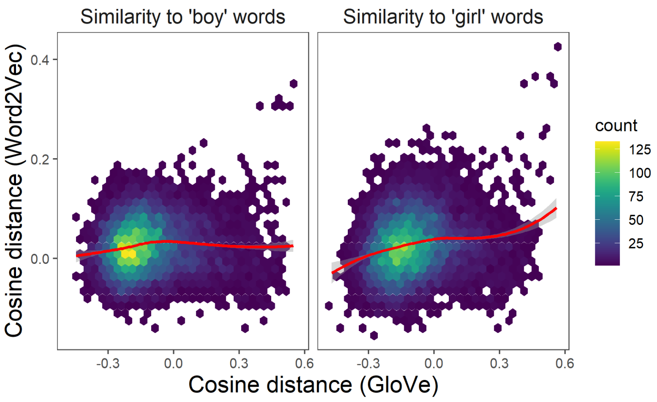

Figure 1: Hexbin plot showing association between gender valence of individual words vectors in Word2Vec and GloVe. Red lines are LOESS smoothers with 95 percent confidence intervals.

To get a rough sense of whether both models are similarly capturing semantic information related to gender, we can plot the average cosine distance to the anchor words from both models on opposing axes (Figure 1). In both similarity to the 'girl' anchor words and similarity to the 'boy' anchor words, the models show a slight positive correlation where the density of points is highest; however, this relationship is not discernible for words with more extreme values of similarity to the anchor words. In addition, a positive correlation between similarity metrics from the two models is more appreciable when measuring individual words' similarity to the 'girl' anchor words (right panel of Figure 1) compared to the 'boy' anchor words.

## Analysis 2: Correspondence with human ratings of word gender valence

Are word embedding models capturing an aspect of individual words' gender valence that corresponds with human intuitions about how gendered those words are? To examine this question, I compute (as in the previous section) the average cosine distance between each word vector in the corpus and the word vectors in the 'boy' anchor words, and subtract this quantity from the average cosine distance between the word vector and the vectors in the 'girl' anchor words. In essence, this quantity roughly captures how much *more* similar a word type is to the 'girl' concept than it is to the 'boy' concept.

To examine whether this quantity captures information about words that accords with human intuitions about how "boyish" or "girlish" a word is, I use gender judgments of over 2,000 word types by human raters from Amazon Mechanical Turk, originally collected by Lewis et al. (2020). Each word was rated by approximately 7 participants, and ratings fall on a 1-5 scale, with 1 indicating that a word is as male-typed as possible and 5 indicating that a word is as female-typed as possible. The number of words for which this dataset contains human ratings is far less than the number of word types present in CHILDES, so a correlation between the word-vector similarity quantity previously mentioned and the human ratings is only possible for a small subset of the CHILDES word types.

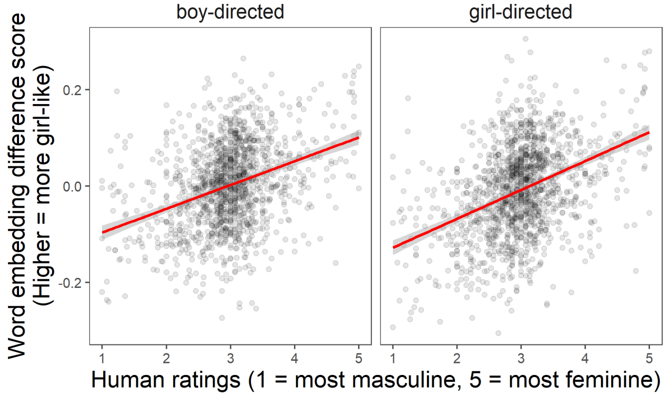

Figure 2: Scatter plot showing individual words gender valence, as determined by human raters (horizontal axis) against word embedding difference score (vertical axis). Lines are best linear fits with associated 95 percent confidence intervals.

Additionally, for this analysis, two separate Word2Vec models were trained: one on speech from CHILDES that was directed at male children, and another on speech from CHILDES directed at female children, to examine if communication to children of different genders shows more or less correspondence with human intuitions about gender valence. Figure **??** shows how these two models, in addition to the one trained on all child-directed utterances (regardless of the child's gender), represent the gender valence of 45 nouns and verbs selected from the MacArthur-Bates Communicative Development Inventory Short Form (Fenson et al., 2000), a language development assessment used with children between the ages of 16 and 30 months[1]. One noticeable pattern from Figure **??** is that word vectors with higher male-valence (i.e., closer to the 'boy' anchor words than the 'girl' anchor words) at least in the model trained on all of the child-directed utterances tend to be modes of transportation, while more female-valenced words tend to be articles of clothing and jewelry.

The results of the word vector correlations with human ratings are displayed in Figure 2. Word embedding models trained either on only speech to girls or speech to boys both capture information about individual words' gender valence that shows a reliable positive association with human ratings of a word's gender valence ($r_{\text{girl-directed speech}} = 0.4001$; $r_{\text{boy-directed speech}} = 0.3514$). The correlations between human judgments and word embedding difference scores underscore two points. First, the gendered nature of certain words, as intuited by human raters, is discernible in the distributional semantics of language directed to children. Second, according to the (very coarse) correlational measure described above, this gender valence of individual words in child-directed speech does not look radically different between speech directed to boys vs. girls, though other analytic strategies could potentially still detect differences in how gendered language is communicated to boys vs. girls, if such differences exist.

## Analysis 3

Does child-directed speech in the CHILDES corpus contain specific gender stereotypes? In Analysis, 3, I examined whether word vector representations derived from CHILDES encode semantic information that encapsulates three particular stereotypes which have been the subject of prior research in social psychology: (1) Female as home-oriented, male as work-oriented; (2) Female as good, male as bad; and (3) Female as oriented towards reading and language, and male as math-oriented. To assess the extent to which these stereotypes might surface in the distributional semantics of the CHILDES corpus, the Word Embedding Association Test [WEAT; Caliskan et al. (2017)] was conducted. The goal of the WEAT is to quantify stereotypical biases in large bodies of text in a manner analogous to how the Implicit Association Test quantifies implicit bias in people (Greenwald, McGhee, & Schwartz, 1998).

Defined more precisely, as in Caliskan (2017), we can consider two equally-sized sets of anchor words $X$ and $Y$ (for example, the sets corresponding to the anchor words for 'boy' and 'girl') and two sets of attribute words $A$ and $B$ (for example, sets of words corresponding to the attributes 'good at reading' and 'good at math'). The effect size generated by the WEAT is

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std dev}_{w \in X \cup Y} s(w, A, B)}$$

where

[1] Nouns and verbs which did not refer to food or animals were chosen for this visualization

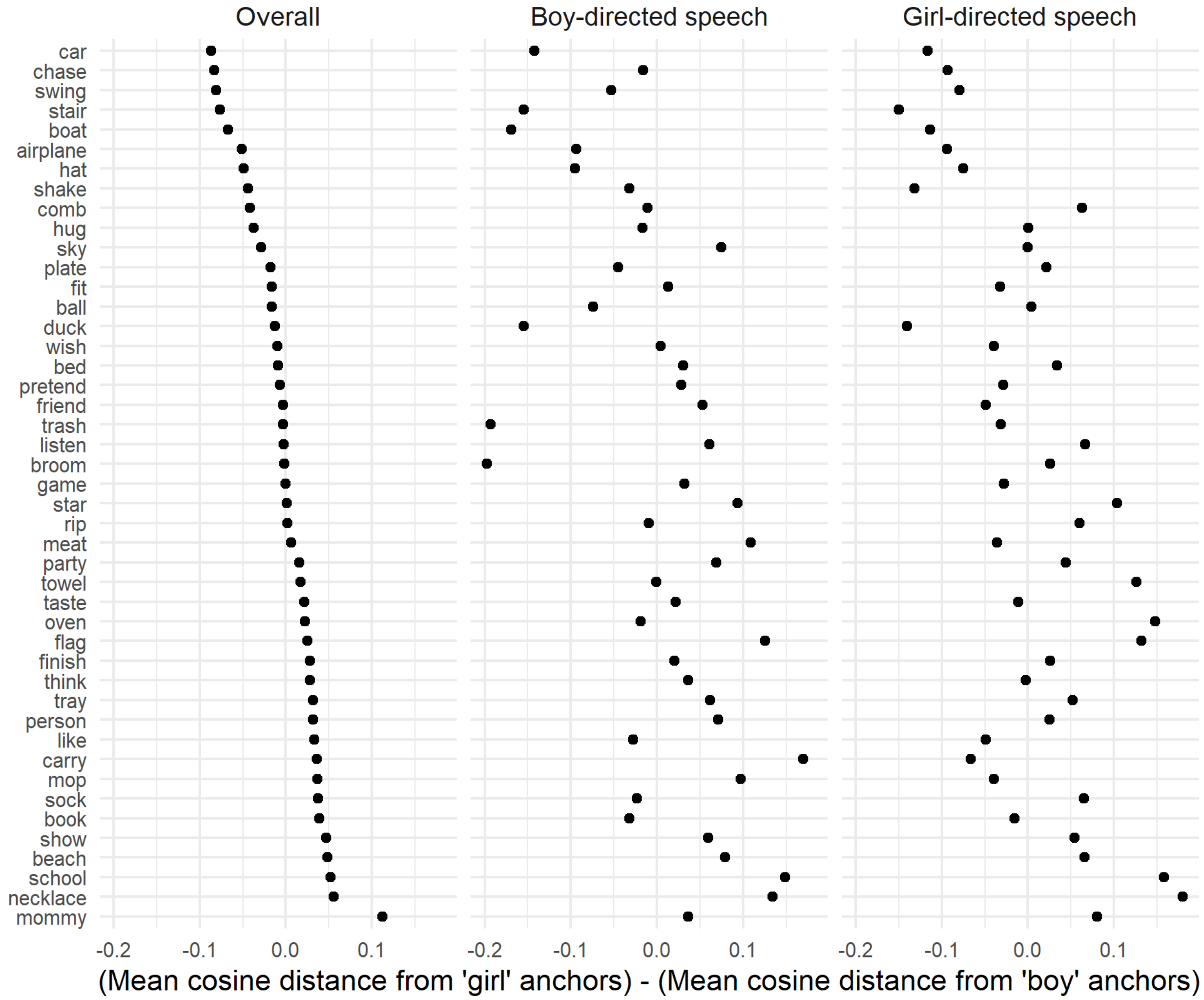

Figure 3: Gendered nature of word vector representations of 45 words from the MacArthur-Bates Communicative Development Inventory Short Form. Horizontal axis represents, for each word on the vertical axis, the difference between the mean cosine similarity of that word to the set of 'girl' anchors and that same word's mean cosine similarity to the set of 'boy' anchors.

Table 1: Words for each attribute set.

| Attribute | Words |
|---|---|
| Home | {family, children, home, cousin, parent, wedding} |
| Work | {job, work, money, office, business, desk} |
| Language | {book, read, write, letter, spell, story} |
| Math | {number, count, sort, size, shape, different} |
| Good | {good, happy, gift, sunshine, heaven} |
| Bad | {bad, awful, sick, trouble, hurt} |

$$s(w,A,B) = \text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})$$

For each of the three previously mentioned stereotypes, a WEAT score was computed using 3 Word2Vec embeddings: one trained only on boy-directed speech, one directed only on girl-directed speech, and one trained on both. The words in each attribute set are displayed in Table 1.

Results from the WEAT test are displayed in Figure 4. Ninety-five percent confidence intervals were generated for WEAT effect size estimates by randomly shuffling which target words corresponded with the “boy” concept and which with the “girl” concept, and subsequently computing a WEAT effect size with the shuffled word labels 1,000 times to obtain an empirical null distribution (following the strategy used in Charlesworth et al., 2020). For the Female-Good/Male-Bad stereotype, WEAT computed on the CHILDES data showed

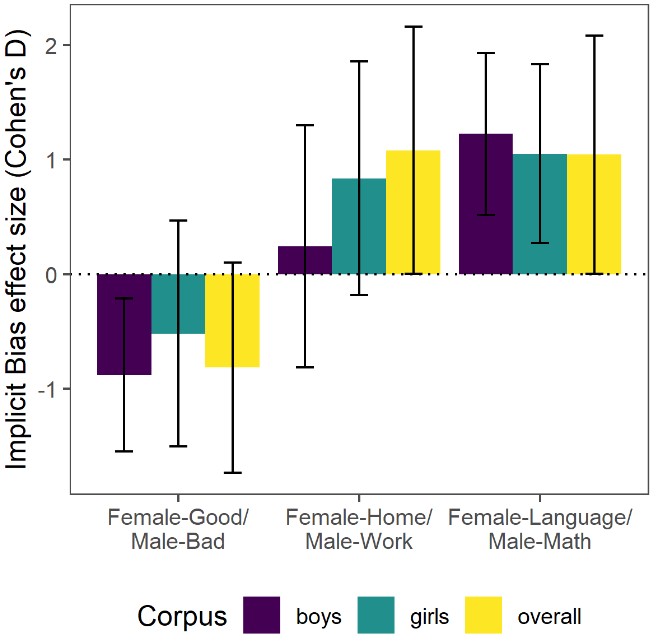

Figure 4: WEAT effect sizes for each stereotype. Error bars are 95 percent confidence intervals on effect size estimates, generated by randomly shuffling target word vector labels and simulating the WEAT effect size 1000 times.

an effect size in the opposite direction from what is considered stereotypical (Cvencek, Meltzoff, & Greenwald, 2011). However, only the effect size generated from speech directed to boys had an effect size estimate whose 95% confidence interval did not include zero. The Female-Home/Male-Work stereotype only showed an effect whose 95% interval did not include zero with a word embedding model trained on speech to both boys and girls; the boy-directed and girl-directed corpora yielded smaller and less robust effects, though all were in the stereotypically-expected direction. The Female-Reading/Male-Math stereotype showed, overall, the largest effect sizes across all three training sets, with all of the three having confidence intervals that did not include 0.

As demonstrated in Figure 4, the results from WEAT on the three CHILDES corpora (all child-directed speech, boy-directed speech, and girl-directed speech) are somewhat mixed. Many of the WEAT calculations generated relatively large effect sizes, but large variance in the empirical null distribution of effect sizes, resulting in wide confidence intervals on the effect size estimates, constrains the interpretation of above data. Nonetheless, some gender stereotypes are discernible in the distributional semantics of the CHILDES corpus, particularly the stereotype that associates boys with math and girls with reading/language.

## Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

## References

10 Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, *45*(1), 106–125.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics*, *21*(1), 95–116.

Gelman, S. A., Ware, E. A., & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive Psychology*, *61*(3), 273–301.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Halim, M. L., & Ruble, D. (2010). Gender identity and stereotyping in early and middle childhood. In *Handbook of gender research in psychology* (pp. 495–525). Springer.

Lewis, M., Borkenhagen, M. C., Converse, E., Lupyan, G., & Seidenberg, M. S. (2020). What might books be teaching young children about gender?

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028.

MacWhinney, B. (2009). The CHILDES project. *Tools for Analyzing Talk–Electronic Edition*, *2*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781*.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Retrieved from `http://www.aclweb.org/anthology/D14-1162`

Ruble, D. N., Martin, C. L., & Berenbaum, S. A. (2006). Gender development.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941.