COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13146

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13146

The Emergence of Gender Associations in Child Language Development

Ben Prystawski,[a] Erin Grant,[b] Aida Nematzadeh,[c] Spike W. S. Lee,[d,e] Suzanne Stevenson,[f] Yang Xu[f,g]

[a]*Department of Psychology, Stanford University*
[b]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley*
[c]*DeepMind*
[d]*Department of Psychology, University of Toronto*
[e]*Rotman School of Management, University of Toronto*
[f]*Department of Computer Science, University of Toronto*
[g]*Cognitive Science Program, University of Toronto*

Received 8 April 2021; received in revised form 9 April 2022; accepted 14 April 2022

Abstract

Gender associations have been a long-standing research topic in psychological and social sciences. Although it is known that children learn aspects of gender associations at a young age, it is not well understood how they might emerge through the course of development. We investigate whether gender associations, such as the association of dresses with women and bulldozers with men, are reflected in the linguistic communication of young children from ages 1–5. Drawing on recent methods from machine learning, we use word embeddings derived from large text corpora including news articles and web pages as a proxy for gender associations in society, and we compare those with the gender associations of words uttered by caretakers and children in children's linguistic environment. We quantify gender associations in childhood language through *gender probability*, which measures the extent to which word usage frequencies in speech to and by girls and boys are gender-skewed. By analyzing 4,875 natural conversations between children and their caretakers in North America, we find that frequency patterns in word usage of both caretakers and children correlate strongly with the gender associations captured in word embeddings through the course of development. We discover that these correlations diminish from the 1970s to the 1990s. Our work suggests that early linguistic communication and social changes may jointly contribute to the formation of gender associations in childhood.

Keywords: Child speech; Gender association; Language and gender; Language development; Social change; Word embedding

Correspondence should be sent to Ben Prystawski, Department of Psychology, 450 Jane Stanford Way, Building 420 Stanford, California 94305 USA. E-mail: benpry@stanford.edu

1. Introduction

Gender associations have been a long-standing topic in psychological and social sciences (Ellemers, 2018). It is believed that children at an early age can make certain gendered associations, such as that trucks are for boys and dolls are for girls (e.g., Fagot, Leinbach, & O'boyle, 1992; Meyer & Gelman, 2016; Raag, 1999). Less understood is how gender associations might emerge through time—over the course of child development—and how these developmental patterns might shift over history as societal attitudes toward women and men change. We investigate the temporal emergence of gender associations in children's linguistic input and output during development and through historical periods via a large-scale analysis of child–caretaker speech corpora that is informed by quantitative tools for capturing associations in machine learning.

Previous research has suggested that language and gender are intricately related in adults and children. Work in sociolinguistics and psychology has documented gender differences in discourse, speech style, language use, and language development (Bamman et al., 2014; Coates, 2015; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Hall & Bucholtz, 2012; Ehrlich, Holmes, & Meyerhoff, 2014; Lakoff, 1973; Laserna, Seih, & Pennebaker, 2014; Lovas, 2011; Newman, Groom, Handelman, & Pennebaker, 2008; von der Malsburg, Poppels, & Levy, 2020). Lakoff (1973) provided an anecdotal account of qualitative differences between how men and women typically speak, noting, for instance, that women are more likely to use hedges such as "John is here, isn't he?" than men. The underlying assumption here is that by analyzing which words and grammatical constructions are more likely to be used by women or men, one can learn about the gender norms of society. If a word is said by men more often than by women, it is likely related to male gender roles. More recent studies have applied a similar but quantitative methodology and reported further gender differences in language use. For instance, men tend to use the possessive *my* in mentioning their spouse or partner, more frequently than women (Schwartz et al., 2013). Men and women also differ slightly, yet significantly, in their ratings of pleasantness, imagery, and familiarity of words (Bellezza, Greenwald, & Banaji, 1986). Related work in psychology has indicated that there are gender differences in child language development. In a meta-analysis, Leaper, Anderson, and Sanders (1998) found differences in parent-to-child speech governed by both the gender of the parent and the gender of the child, such as that mothers use more supportive language when speaking to daughters than to sons and that fathers use more directive language than mothers overall. Further evidence for gender associations in the linguistic environment of children has been observed in television programs. These studies focus on how language use differs as a function of the recipient, which elucidates a possible mechanism for the transmission of gender associations. If girls are more likely to hear language about dresses and dolls while boys are more likely to hear about trucks and sports, differences in linguistic input could socialize children into traditionally female and male gender roles.

Existing work has also analyzed the content of cartoons and children's television shows and reported that language used in these shows contains stereotypical gendered associations (Aubrey & Harrison, 2004; Mulac, Bradac, & Mann, 1985). For example, male characters tend to associate more frequently with action verbs and present tense verbs, whereas

female characters tend to associate more with uncertainty verbs and polite terms (Mulac et al., 1985). Here, the researchers analyzed patterns of co-occurrence between characters in media and particular words, which helps to understand whether the language presented to children linking male and female subjects with different concepts might lead children to form beliefs about what behavior is typical or expected of women and men.

A separate line of research has demonstrated that young children learn gender associations at an early age (Martin & Dinella, 2001), and they do so sometimes via implicit cues from their parents (Endendijk et al., 2014). Moreover, children younger than 3 years of age show some ability to make gendered associations and labels (Fagot et al., 1992). As early as the first grade, children already endorse the stereotypical belief that boys are more interested in computer science and engineering than girls (Master, Meltzoff, & Cheryan, 2021). Boys and girls also learn different words earlier and later in childhood, which is likely a result of gendered patterns of play (Frank et al., 2021). These studies have consolidated the view that gender associations are present in childhood, but they are less informative about the precise pattern through which gender associations emerge in early life.

Understanding the long-term cognitive, emotional, and socializing impacts of children's differential exposure to words relies on longitudinal corpora in which the same children are studied over several years. Such corpora are sparse, so we instead focus on quantifying differences in which words boys and girls are exposed to. We explore the emergence of gender associations in child language development by considering the relations of three measures: (a) We measure differences in linguistic input by gender to quantify the gender associations conveyed in the linguistic communication[1] from caretakers' speech; (b) we measure gender differences in linguistic output by gender of the child to understand gender associations conveyed in children's speech; and (c) we quantify patterns of co-occurrence and association within language from the broader community to understand whether the gender associations in society are reflected in linguistic communication during childhood, namely, how they correlate with measures described in (a) and (b).

To operationalize the first two measures, we draw on the CHILDES corpus—one of the largest corpora of child-directed speech (CDS) and child speech (CS). This corpus contains transcripts of naturalistic conversations between caretakers and children, tagged by both child and caretaker gender. We develop a new metric, called *gender probability*, to quantify the relative frequencies of a word's usage in the speech of caretakers and children. We focus on the English-language version of the corpus because other languages either have less data available or have the confound of grammatical gender in adjectives and nouns.

To operationalize the third measure and quantify gender associations at scale in the language use of the broader society, we use word embeddings from machine learning—vector representations of word meaning trained on large text corpora external to the CHILDES corpus—which are known to capture implicit gender associations (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018; Mikolov, Chen, Corrado, & Dean, 2013a; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013b; Pennington, Socher, & Manning, 2014). Common models such as Word2Vec (Mikolov et al., 2013a, 2013b) and GloVe (Pennington et al., 2014) learn to represent a word's meaning as a real-valued vector based on its co-occurrences with other words in sentences. The resulting word vector representations can then capture

similarity relations such as *queen* is similar to *king* (in meaning), as well as simple analogies such as *queen is to king* is analogous to *woman is to man*. Due to these properties, word embeddings have been shown to robustly capture people's implicit gender associations such as nurses as female and engineers as male (e.g., Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Caliskan, Bryson, & Narayanan, 2017; Garg, Schiebinger, Jurafsky, & Zou, 2018; Lewis & Lupyan, 2020). Bolukbasi et al. (2016) also showed that the difference between the pre-trained Word2Vec and GloVe embeddings for *man* and *woman* is approximately equal to the difference between the embeddings for *computer programmer* and *homemaker*: $\vec{\text{man}} - \vec{\text{woman}} \approx \vec{\text{computerprogrammer}} - \vec{\text{homemaker}}$. This vector algebra can be interpreted as the analogy "Man is to woman as computer programmer is to homemaker," a statement that expresses a stereotypical gender association in occupation. More recent work has followed up on these studies by showing that gendered associations in word embeddings are robust across source text corpora and reflected in different languages (Lewis & Lupyan, 2020). It has also been shown that gender associations in word embeddings accurately predict behavioral data in psychological implicit association tests (Caliskan et al., 2017) and public records of gender imbalances across professions (Garg et al., 2018), indicating that they reliably reflect gender associations in society at large.

Latent semantic analysis, an alternative method for constructing vector representations of word meanings, has also been shown to capture gender associations in role words (Lenton, Sedikides, & Bruder, 2009), though its capacity to reflect the gender associations of general society has been less thoroughly documented. It may also be possible to measure gender associations via behavioral experiments. For example, we could either ask people to directly rate whether and how a word is gendered or measure implicit associations (e.g., Greenwald, Poehlman, Uhlmann, & Banaji, 2009). The former method could miss subtle associations that depend on context. For instance, people might rate a word like "doll" as non-gendered but still use it disproportionately when speaking about girls. The latter method could capture implicit associations, but is difficult to scale up to large numbers of words. Word embeddings provide the benefits of capturing implicit associations and scaling to the broad lexicon.

We first hypothesize that linguistic communication might contribute to the early formation of gender associations. In particular, we expect the word frequencies in children's linguistic input from caretakers to reflect broad gender associations, and given the intimate relations between caretakers' CDS and the speech produced by children (e.g., Singh, Nestor, Parikh, & Yull, 2009; Shneidman & Goldin-Meadow, 2012; Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013), we expect the word frequencies in CS itself to reflect gender associations (see Fig. 1 for an illustration). We next hypothesize that the strength of gender associations in speech has decreased over the recent decades as a reflection of the social changes related to gender during the time period of our study, including a greater entry of women into the workplace and more egalitarian public attitudes (Donnelly et al., 2016; Eagly, Nater, Miller, Kaufmann, & Sczesny, 2019).

To evaluate our hypotheses, we examine the correlation between gender probability in caretaker–child speech (from CHILDES) and word embedding associations in vectors trained on large corpora of general language. This enables us to estimate how much of the variance in how frequently words are said to and by girls and boys can be explained by gender

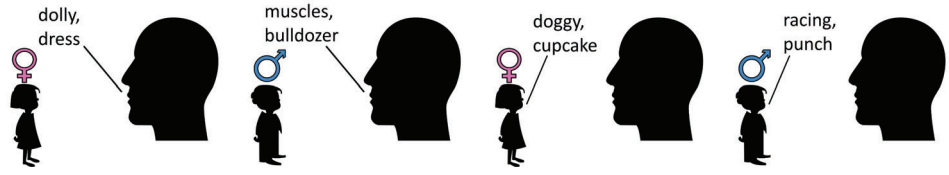

Fig. 1. An illustration of the hypothesis on gendered speech between young children and their caretakers.

associations present in general language use. In other words, we want to understand the extent to which the gender associations in society are reflected in children's linguistic input and output and when they emerge through language development. We examine these issues focusing on English-speaking children from age 1 to age 5. We also investigate how the strength of gender associations in child language development has changed over history, for the period 1970–2000. Our approach helps to address a series of probing questions regarding gender association in CS and caretaker speech, including its emergence through development, its historical trends, and its potential social roots.

Our work differs from and extends the study by Charlesworth, Yang, Mann, Kurdi, and Banaji (2021) that demonstrates the presence of gender biases in corpora of adult and CS. That study uses word embeddings trained directly on different corpora to test for gender biases in speech, books, and audiovisual media for adults and children. While their work focuses on demonstrating the presence of gender biases in childhood language and media, we investigate how gender associations develop through time: how they emerge in childhood and change over historical periods. We use word embeddings that are pre-trained on large text corpora of general language use (i.e., independent to the data on child–caretaker speech) to approximate gender associations in society. We analyze the degree to which children's linguistic environment reflects these associations and how it varies with respect to variables such as child age and historical period by estimating gender probability independently for each age or historical period. Neither of these analyses was examined in Charlesworth et al. (2021), since it generally requires a large amount of data to train reliable word embeddings, and hence, subdividing the corpus by variables like age could yield data too sparse to produce trustworthy results. For these purposes, our methodology has the advantage of not requiring us to train word embeddings individually on each subset of the corpus. Our methodology also allows us to leverage the gendered information contained in more total words: any word included in the vocabulary of the word embeddings can be used rather than only those that were pre-selected to apply the Word Embedding Association Test (WEAT) on.

2. Methods

2.1. Data and code availability

We used public databases for our analyses. Specifically, CHILDES data are publicly available at https://childes.talkbank.org/. The Santa Barbara Corpus is available at

https://www.linguistics.ucsb.edu/research/santa-barbara-corpus. The Switchboard corpus is available at http://compprag.christopherpotts.net/swda.html and was accessed through the ConvoKit Python library (Chang et al., 2020). Pre-trained word embeddings are available at https://github.com/mmihaltz/word2vec-GoogleNews-vectors (Word2Vec), https://nlp.stanford.edu/projects/glove/(GloVe), and https://fasttext.cc/docs/en/crawl-vectors.html (fastText).

Code for reproducing our analyses is deposited at the following link:
https://osf.io/635em/.

2.2. Pre-trained word embeddings

We used three commonly used sets of pre-trained word embeddings: the Word2Vec embeddings trained on the Google News corpus (Mikolov et al., 2013a, 2013b), the GloVe embeddings trained on the Common Crawl corpus (Pennington et al., 2014), and the fastText English embeddings trained on the Common Crawl and Wikipedia corpora (Grave et al., 2018).

The corpora on which these word embeddings were trained are very large. The Google News corpus contains approximately 100 billion tokens of English news articles from various publications. The Common Crawl corpus contains approximately 840 billion tokens, which were scraped automatically from the internet. The Wikipedia corpus contains the text of the Wikipedia dump from September 2017, on the order of billions of tokens. While these corpora generally reflect text on the internet or written sources, they are commonly taken to represent general language use and word associations present in everyday language. Additionally, Caliskan et al. (2017) found that Word2Vec and GloVe embeddings trained on these corpora showed similar biases and associations to those that the Implicit Association Test (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Nosek, Banaji, & Greenwald, 2002) reveals in humans.

2.3. Embedding-based quantification for gender associations

To quantify people's gender associations using word embeddings, we consider two representative formulations of gender association tests from work in machine learning: the WEAT (Caliskan et al., 2017) and gender Subspace Projection (PROJ) (Bolukbasi et al., 2016).

The WEAT is a common procedure for measuring associations between word embeddings (Caliskan et al., 2017). In this test, the effect size of a given word's association is the difference between the mean cosine similarities between the word's vector and those in two sets of attribute words:

$$s(w, F, M) = \frac{\text{mean}_{f \in F} \cos(\vec{w}, \vec{f}) - \text{mean}_{m \in M} \cos(\vec{w}, \vec{m})}{\text{std\_dev}_{x \in M \cup F} \cos(\vec{w}, \vec{x})} \quad (1)$$

Here F and M denote the two sets of attribute words—gender terms in this case, w denotes the word in question, and $\vec{x}$ denotes the vector associated with word x in the joint set of the attribute words. p-Values from this test are calculated using a permutation test. We used the same sets of male and female terms used in the original WEAT formulated by Caliskan et al. (2017), where they found that this procedure applied to groups of target words can closely

reproduce gender associations from implicit association tests in psychology (Greenwald et al., 2009; Nosek et al., 2002).

PROJ (Bolukbasi et al., 2016) offers an alternative method to characterize gender association in word embeddings. This method first identifies a gender subspace by performing principal component analysis (PCA) on the vector differences between pairs of words that differ in gender, such as (woman, man) and (mother, father). The first principal component explains approximately 60% of the variance (Bolukbasi et al., 2016) and it is taken to be the axis of the one-dimensional gender subspace. An individual word's gender association is then quantified by projecting that word onto the gender subspace: $\text{proj}_G(\vec{v}) = \frac{\vec{v}\cdot\vec{b}}{|\vec{b}|}$. Here G is the gender subspace, $\vec{b}$ is the basis vector for the gender subspace identified via PCA, and $\vec{v}$ is the vector of the word in question. This returns a number between $-|\vec{v}|$ and $|\vec{v}|$, where positive numbers reflect more female-associated words and negative numbers correspond to male-associated words.

Ethayarajh, Duvenaud, and Hirst (2019) show that WEAT may overestimate the strength of associations and note that PROJ is not subject to the tendency to overestimate bias. Here we consider both WEAT and PROJ to ensure that our computational analysis is robust to methodological choices.

2.4. Quantification of gender probability in speech

We measure the gender association of a word's usage in natural speech by a probability metric, termed *gender probability*, that is obtained independently to the gender associations quantified by WEAT or PROJ. We estimate this metric empirically based on how often a word is used to or by children from one gender compared to the baseline frequency of words being said to children of that gender in the corpus. This metric can be used to quantify the degree of gender association either from the caretakers' perspective (i.e., CDS) or children's perspective (i.e., CS). Formally, we define the gender probability of a word via Bayes' rule:

$$p(g|w) \propto p(w|g)p(g) \tag{2}$$

Here g stands for gender $g \in \{f, m\}$, where f denotes female and m denotes male. For the scope of this work, we assumed that gender variable g is binary and distinguishes between female and male. This is how gender is labeled in the corpora that make up the CHILDES corpus, although there exists work that took more nuanced approaches to modeling gender in text such as from social media (Bamman et al., 2014). w represents a target word in question. We assume a uniform prior on the gender of the interlocutor (i.e., $p(f) = p(m) = .5$). To ensure that the results we report are not artifacts of our decision to use a uniform prior, we analyzed 10,000 random subsamples of the corpus in which we enforced equal representation of boys and girls and ages 1–5. The number of tokens in each category of each subsample was the total number of tokens in the corpus divided by the number of categories. For example, in the yearly analysis, we had one category for each combination of age (1–5) and gender (m, f), for a total of 10 categories per speech type (CS, CDS). This ensures that categories are represented equally and subsamples have the same number of tokens as the original corpus. We estimate the likelihood of a word w for a specific gender as the relative frequency that w

is associated with that gender in context. For example, the female likelihood is:

$$p(w|f) = \frac{c(f, w)}{c(f)} \tag{3}$$

Here $c(g, w)$, where $g = f$ is the number of times word w is said to or by children with gender g, and $c(g)$ is the total count of all words said to or by children with gender g in the corpus. It then follows that the (female) posterior probability is:

$$p(f|w) = \frac{\frac{c(f,w)}{c(f)}}{\frac{c(f,w)}{c(f)} + \frac{c(m,w)}{c(m)}} \tag{4}$$

This metric takes into account the base rate difference in words said to or by female versus male children. For instance, if a word w appears frequently in group f, the relative frequency of w could be quite low in group f if there are many words said to group f and few words said to group m overall (i.e., even if $c(f, w) > c(m, w)$, $p(f|w)$ could be still lower than $p(m|w)$ in principle). Under this metric, a word said exclusively to or by girls would have a female gender probability of 1, and a word said exclusively to or by boys would have a female gender probability of 0. We have also considered alternative approaches to quantify gender association in speech such as log odds ratio and our results are robust to this variation (see Supporting Information for details).

In our analyses, we calculate gender probability separately for caretakers and children with respect to children's gender. From the caretakers' perspective, we measure the gender probability of words said to children considering children as listeners (i.e., gender association in CDS). From the children's perspective, we measure the gender probability of words said by children toward caretakers considering children as speakers (i.e., gender association in CS).

2.5. Corpus of adult speech

To estimate the strength of gender associations in adult–adult speech as a comparative dataset to caretaker–child speech, we used the Santa Barbara corpus (Du Bois, Chafe, Meyer, Thompson, & Martey, 2000). This corpus consists of 340,860 tokens of naturalistic speech representing 60 conversations between adults. Speakers in the corpus are not tagged by gender, but their names are provided. We tagged speakers by gender by checking their name in lists of predominantly male and predominantly female names (Kantrowitz, 1991). Speakers whose names did not occur in either list were excluded from the analysis.

We also considered the Switchboard corpus as a secondary resource for adult–adult speech (Godfrey, Holliman, & McDaniel, 1992). This corpus contains 1,531,972 tokens representing 1,155 telephone conversations between adults. Speakers are tagged by sex. Analyses of the strength of gender associations in the Santa Barbara and Switchboard corpora can be found in the Supporting Information.

2.6. Hall corpus

Although the majority of CHILDES subcorpora do not contain fine-grained demographic information (see Supporting Information for more details), the Hall corpus is a special and

large subcorpus of naturalistic speech from families around the 1980s, which is explicitly tagged both by socioeconomic class and race (Hall & Tirre, 1979; Hall, Nagy, & Nottenburg, 1981; Hall, Nagy, Linn, & Bruce, 1984). This corpus includes speech from 36 children age 4 and contains a total of 1,275,572 tokens. Families fall into one of four categories defined by the original studies from the Hall corpus (Hall & Tirre, 1979; Hall et al., 1981, 1984): White working-class (WC), Black WC, White middle-class (MC), and Black MC. In those studies, WC families refer to households where children attended exclusively public schools, and MC families refer to households where children attended exclusively private schools. We use the Hall corpus to analyze differences in the strength of gender associations by social class and race.

2.7. Bootstrapped hypothesis testing

We used bootstrapping to test for significance in our correlation-based results for CS, CDS, and adult speech. We defined both one-sample and two-sample hypothesis tests. In the one-sample hypothesis test, the test statistic is the correlation strength ρ, and we test for a meaningful difference between ρ and 0. This is achieved using the standard method for bootstrapped hypothesis testing: subtracting the mean from the empirical distribution of correlation strengths to create a null distribution, then measuring the proportion of samples from the null distribution whose absolute value exceeds that of the measured test statistic. Our test is therefore two-tailed. In the two-sample hypothesis test, we use the same general procedure, except our test statistic is the difference between the two categories (denoted as d). Since we make many comparisons across different speech types, word embeddings, and association metrics, we report adjusted p-values using Bonferroni correction for multiple comparisons. Our bootstrapping method involves resampling the entire corpus and repeating the correlation analysis with the bootstrapped subsamples. To the best of our knowledge, there is no method to aggregate different samples taken this way for correlation (e.g., into a mixed-effects model), so we report separate tests for each combination of association metric and word embedding type. We created 10,000 bootstrapped subsamples of the corpus. Across bootstrapped subsamples, there were on average 4,557 words of CDS and 2,869 words of CS above the frequency threshold of 20. An average of 2,609 of these words are shared between CS and CDS. This reflects the fact that there are more words of CDS than CS overall in CHILDES, but the words that appear repeatedly in CS tend to also appear in CDS.

3. Results

3.1. Gendered speech in children and caretakers

We evaluate our hypotheses using data from natural speech between English-speaking children and caretakers in North America. We used CHILDES (MacWhinney, 2014), the largest publicly accessible inventory of child-caretaker speech commonly used in psychology and cognitive science. We pooled data across children and caretakers from the North American section of CHILDES, filtering to include only speech from normally developing children

(based on corpus information in CHILDES about language disorders) and in naturalistic conditions. The metacorpus we worked with includes 9,417,152 tokens and mostly covers the ages of 1–6. Each conversation in the corpus included exactly one child, and the speakers were labeled by gender. We labeled a word as said to or by a child of a particular gender based on the gender of the child in the conversation and the participant saying the word. There were a total of 4,875 children in the subset of the corpus we analyzed, with 2,169 boys and 2,706 girls. The percentage of words that were said to girls for each age is as follows: Age 1–51%, Age 2–50%, Age 3–45%, Age 4–38%, Age 5–42%, Age 6–13%. To investigate the linguistic behavior of preschool children (and considering the imbalance in the data from age 6), we focused our analyses on ages 1–5. In analyzing the data from this corpus, we also excluded people's names and proper nouns because lexical items such as names have dedicated functions for differentiating gender (e.g., *Linda* is typically a woman's name and *Michael* is typically a man's name), as well as explicitly gendered words like "mom," and "husband." A full list of excluded words can be found in the Supporting Information.

We begin by examining whether the gender probabilities of the words said to and by children in the CHILDES corpus correlate with the gender associations in everyday language (e.g., from news and online media) obtained independently in the word embeddings. To ensure robustness of the analysis, we considered three types of word embeddings commonly used in machine learning: Word2Vec (Mikolov et al., 2013a, 2013b), GloVe (Pennington et al., 2014), and fastText (Grave et al., 2018). For each embedding method, we also applied two established tests for estimating the gender associations of words in the lexicon: WEAT and PROJ.

Fig. 2 summarizes the strength of overall correlations between gender probabilities of words in child–caretaker speech for ages 1–5 and gender associations of the same words in word embeddings from WEAT and PROJ tests. For this analysis, we kept common words that occurred at least 20 times in both CDS and CS in the given bootstrapped instance of the CHILDES corpus that were also represented in the vocabulary of the pre-trained word embeddings. This resulted in over 1000 words used to compute the correlation in all bootstrapped instances of the corpus. We found a significant and robust Pearson correlation between gender probability and gender associations in word embeddings, with $p < .01$ in all of the 12 tests that we performed: 2 (child-directed vs. child speech) $\times$ 2 (WEAT vs. PROJ) $\times$ 3 (Word2Vec, Glove, fastText). p-Values are Bonferroni-corrected to account for the number of hypotheses tested. Importantly, we observed that the correlation strengths found in children's speech are similar to those found in caretakers' speech.

We also compared these findings to gender associations in adult–adult speech as a control set, which we measured in the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000) and the Switchboard corpus of telephone speech (Godfrey et al., 1992). We used the same number of bootstrapped iterations and the same frequency threshold for adult speech as we did for CS in the CHILDES corpus. The correlation strengths in this adult speech corpus vary widely by measure and embedding type, but the maximal Pearson correlation does not exceed .31 (Switchboard corpus, GloVe embeddings, WEAT: $\rho = .31$, $p < .01$). See Supporting Information for more details of the analyses of adult–adult speech. These results show that gender associations in speech to and by children can be similarly strong to those in speech between adults.

Fig. 2. Correlations between gender probability (in child-directed speech, or CDS, and child speech, or CS) and gender association in word embeddings. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the CHILDES corpus that were balanced by age and gender. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

Fig. 3 shows a subset of words in the CHILDES lexicon to illustrate the similarities and differences between gender probability in children's linguistic environment and gender associations in word embeddings from WEAT. We sampled an equal number of words from each quadrant to illustrate both concordant and discordant cases. In a random sample of words, we would see more words in the top-right and bottom-left quadrants. Words that fall in the bottom-left and top-right quadrants correspond to concordant cases between the two measures. In particular, words of action and strength (cf. Mulac et al., 1985) such as *throw* and *bolt* are found to be consistently male-associated based on our measure and WEAT. Words for female-associated occupations, like *nurse*, and words related to female-oriented toys such as *dolly*, are found to be consistently female-associated. Not all words are concordant between the two measures. For instance, some food words such as *oatmeal* and *pepper* are associated with one gender in their word embedding associations but another in gender probability.

To assess whether the strength of gender associations in CS varies by both caretaker gender and child gender, we compared the average gender association of words across (parent gender, child gender) pairs. Specifically, we computed the average word embedding gender association of words, weighted by frequency, said by a parent of a particular gender to a child of a particular gender. For example, we would count all the words said by mothers to sons and compute a weighted average of their WEAT and PROJ scores, respectively. We found that the gender associations of words differ by the gender of the caretaker as well as the

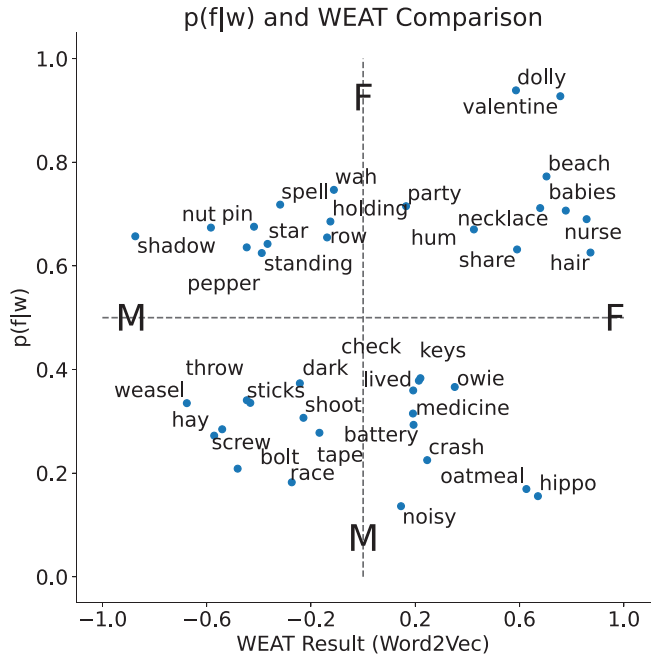

Fig. 3. Word samples that illustrate the measures of gender probability for child–caretaker speech and WEAT for word embeddings on a subsample of words in the CHILDES corpus. The x-position of a point shows its WEAT association: words further to the right are more female-associated in Word2Vec embeddings. The y-position of a point shows its gender probability in CDS. Words further up are said more to girls in the CHILDES corpus.

gender of the child. Fig. 4 shows the mean gender association in CDS for each parent–child gender pair (e.g., mother to son) and each set of word embeddings. The mean association is slightly male in all categories, but the differences between categories reveal an intuitive pattern: speech to boys is more male-associated than speech to girls and speech by fathers is more male-associated than speech by mothers. There is substantial variation between different sets of word embeddings and different measures (WEAT vs. PROJ), but the general pattern holds for all embeddings and measures we considered. The mean association of words said by fathers is more male than words said by mothers, and this effect is statistically significant for all the measures and embeddings ($p < .01$). The mean association of words said to sons is more male than that of words said toward daughters, which is statistically significant for all combinations of association tests and word embeddings ($p < .01$ in all cases). All hypothesis tests were conducted using the two-sample bootstrapping-based method we described. p-Values are Bonferroni-corrected to account for all parent–child pairs, association tests, and word embedding types.

Taken together, these initial findings provide strong evidence that gender associations are reflected in word usage during early child development, both in children's linguistic input and output.

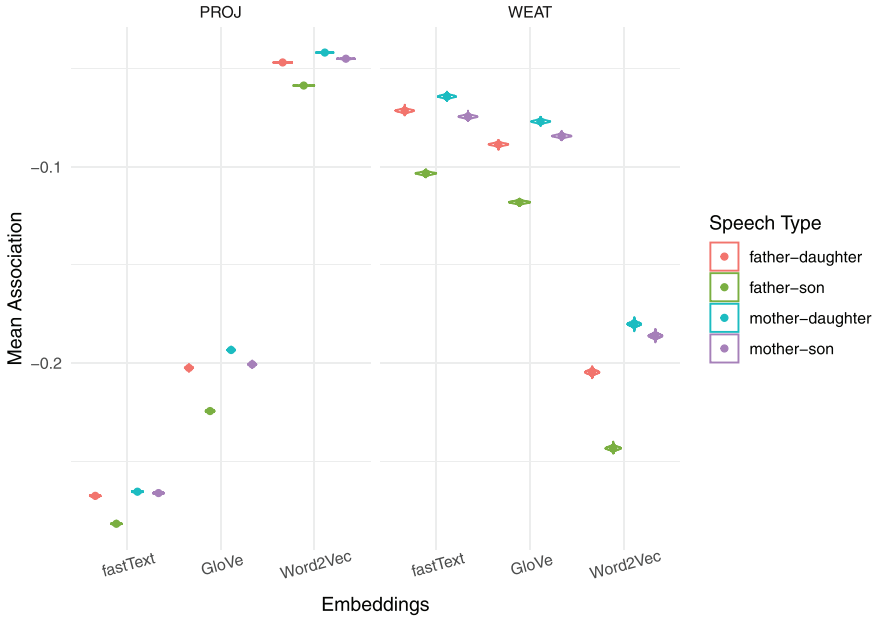

Fig. 4. Average word embedding association in child-directed speech for words said by parents in each parent gender–child gender pair. Higher y-values indicate that words are more female-associated on average. Mean associations were computed by taking the average word embedding association score, using either WEAT or PROJ, weighted by the frequency of the word. Raindrop plots show the density of mean word embedding associations across 10,000 bootstrapped subsamples of the CHILDES corpus.

3.2. *The developmental time course of gendered speech*

To examine the time course of gendered speech, we performed a stratified analysis of the CHILDES corpus by child age. We focused on tracking the emergence of gendered language for children ages 1–5. For each age, we measured correlations between word embedding associations and gender probability for words said at least 20 times by and to children of that age in the CHILDES corpus. We varied the frequency threshold between 10 and 50 and found that the results are robust to this variation. All p-values in this section are Bonferroni-corrected to account for multiple comparisons across age, speech type, association test, and word embedding type.

We computed gender probability for the subset of words in the corpus said to and by children of each age, and then measured the correlation between the gender probabilities and gender associations in word embeddings, showing the extent to which speech to or by children of a particular age reflects the gender associations of society at large. Fig. 5 shows that the correlation strength between word embedding associations and gender probability in CS increases sharply between ages 1 and 2, then levels off with a slight dip at age 4. Correlations are insignificant at age 1 in both CS and CDS for all word embeddings and association metrics. Correlations are mostly significant in subsequent years with a few exceptions. By age 5, correlations are significant at $\alpha = .05$ for all but one combination of association test, speech

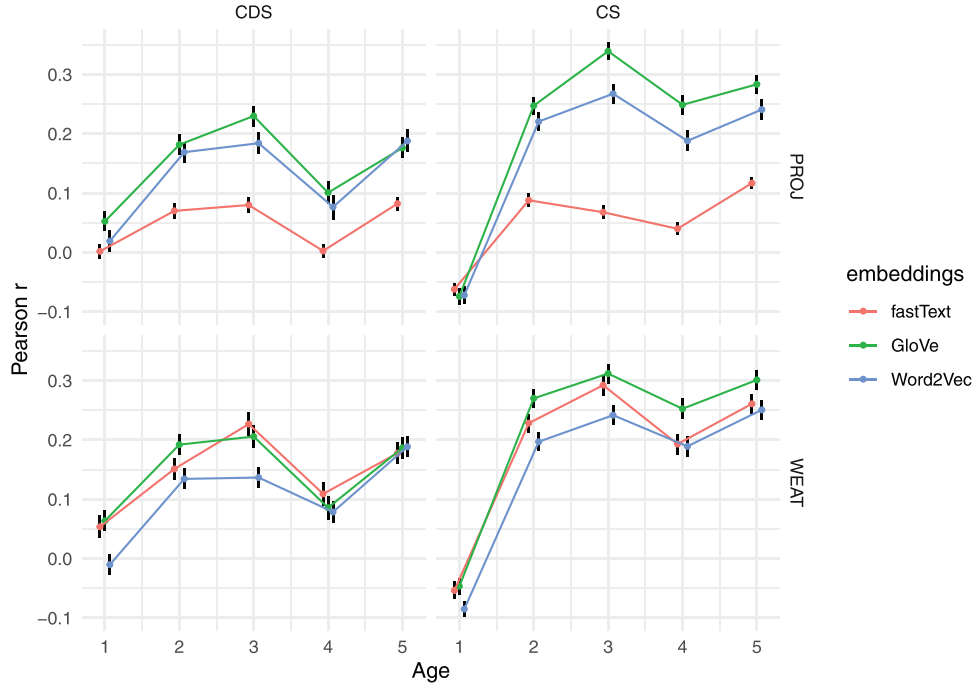

Fig. 5. Developmental time course of correlations between gender probability in child-directed speech (left) and child speech (right) and gender association in word embeddings. Point estimates show the mean correlation across 10,000 bootstrapped subsamples of the CHILDES corpus balanced by age and gender. Error bars denote standard error of the mean.

type, and word embeddings.[2] Full results of hypothesis tests are reported in the Supporting Information. The trend in CDS aligns with the trajectory in CS although it is comparatively flatter. These results suggest not only that gendered word usages in children's linguistic environment correlate with gender associations in word embeddings, but that they also emerge very early in life—around the age of 2 in CS. The alignment in the temporal trajectories between CS and CDS also suggests that linguistic input from caretakers may contribute to the early formation of gender associations.

Fig. 6 visualizes the gender probability of a sample of words in CS in a two-dimensional word embedding space constructed via t-SNE (Maaten & Hinton, 2008). t-SNE is a dimensionality reduction technique that maps high-dimensional vectors like word embeddings to lower-dimensional spaces while preserving some of the closeness relationships of the full-dimensional embeddings.[3] Fig. 6 supplements our quantitative analyses with intuition about the relationship between the positions of word vectors and gender probability. We focused on the 30 words with the highest aggregate gender probability (i.e., said most frequently to girls) and the 30 with the lowest (i.e., said most frequently to boys) among words that occur at least 500 times in the corpus and used gender probabilities from one bootstrapped subsample. Gender probability follows a color scale, where red corresponds to words said more to girls and

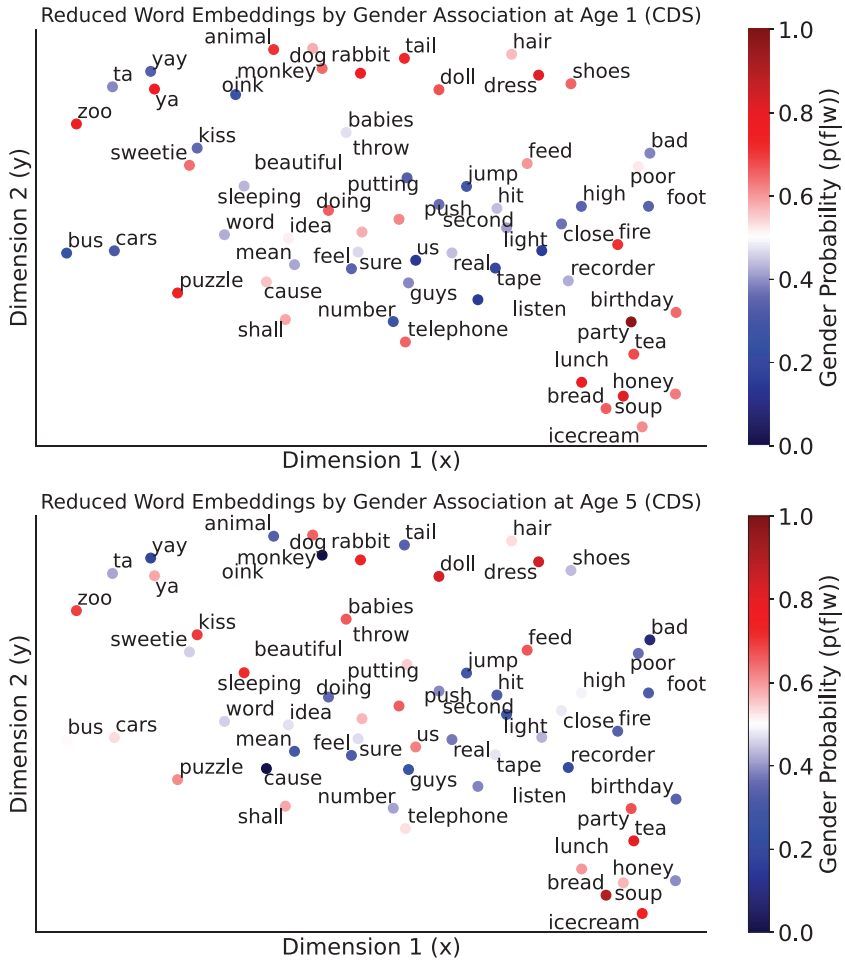

Fig. 6. Visualization of words in child-directed speech that show high (female-oriented) and low (male-oriented) gender probabilities, for age groups 1 and 5 in development. Semantic space is constructed from dimensionality-reduced Word2Vec word embeddings. Colorbar indicates the scale of gender probability, with 1 indicating words exclusively uttered to girls and 0 exclusively to boys.

blue corresponds to words said more to boys. We found clusters of words in the embedding space that share similar gender probabilities belonging to each gender group. For instance, morally valenced words (Luther & Legg Jr, 2010) such as *bad* and action verbs such as *jump* and *push* are said more to boys, while food terms such as *bread* and *soup* are close to each other and said more to girls. We also observed that some of these distinctions between male- and female-oriented words tend to be persistent through the developmental course, illustrated in children at ages 1 and 5. An equivalent version of these figures where points are colored by gender probability in CS can be found in the Supporting Information.

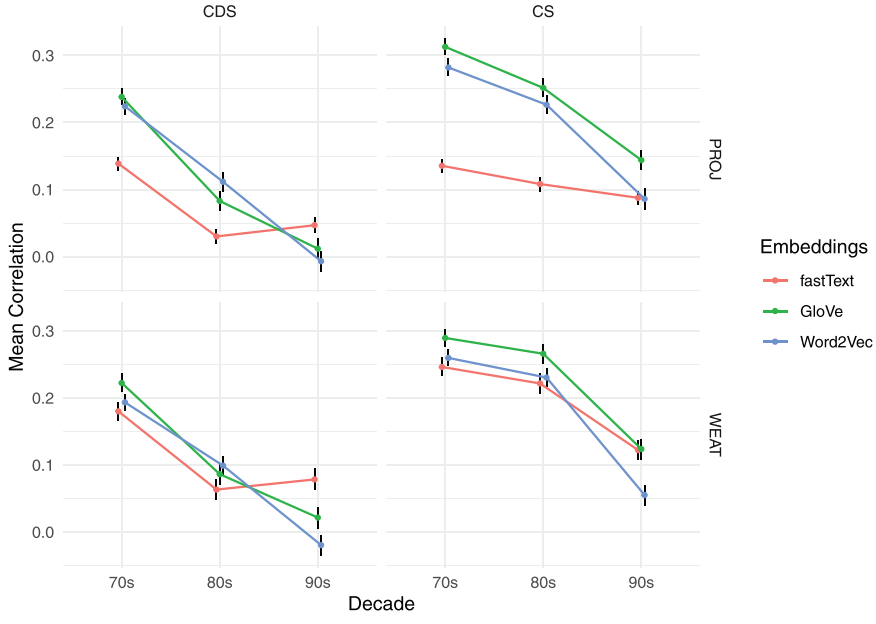

Fig. 7. Correlations of gender probability in each decade of child-directed speech (left) and child speech (right) with gender association in word embeddings based on PROJ (upper row) and WEAT (bottom row) tests. Point estimates show the mean correlation strength across 10,000 bootstrapped samples of the CHILDES corpus and error bars denote standard error of the mean.

3.3. Historical and social influences on gendered speech in childhood

To further investigate gender associations in childhood over history, we performed a stratified analysis of the CHILDES data by decade from the 1970s to the 1990s. We measured the aggregate correlations between gender probability in CDS and CS on the one hand, and word embedding associations for each of the three decades on the other. The results appear in Fig. 7. All p-values in this section are Bonferroni-corrected to account for multiple comparisons across each pair of decades, word embedding type, speech type, and association test. Significance tests on the difference between correlation strengths between the 1970s and 1980s suggested a decrease, though results are not all significant. All of the correlation strengths decreased significantly in CDS for all tests and embeddings. The decrease is not significant in CS for any test or embedding type. Eight of 12 combinations of test, embedding type, and speech type showed a significant decrease in correlation strength between the 1980s and 1990s, while the remaining combinations did not show a significant change. Finally, 11 of 12 combinations showed a significant decrease between the 1970s and 1990s ($p < .05$ in each case). Detailed results of all these tests can be found in the Supporting Information.

These historical results confirm our hypothesis that the strength of the relationship between the gender associations present in society at large and in speech to and by children diminished over time, compatible with societal shifts toward a more egalitarian view. Since we used word embeddings trained on language from the 2010s, it is possible that what changed

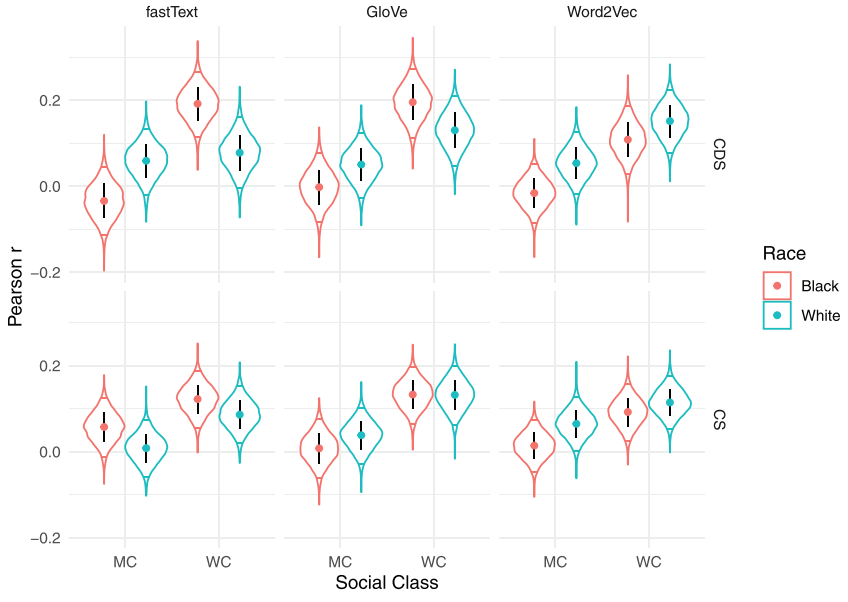

Fig. 8. Correlation strengths of gender probability in child-directed and child speech with word embedding gender associations across socioeconomic classes (working class, or WC, versus middle class, or MC) and racial groups. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the CHILDES corpus that were balanced by gender, race, and socioeconomic class. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean. All plots were created using WEAT.

is not the degree to which gender associations in speech reflect those in broader society in their own times, but how closely they reflect the associations of the time the embeddings were trained. To rule out this possibility, we also analyzed these historical trends using the diachronic HistWords embeddings (Hamilton, Leskovec, & Jurafsky, 2016). This way, we can compare gender probability in speech from a given decade against gender associations in word embeddings trained on text from the same decade. We find the same pattern using diachronic word embeddings: correlations decrease from the 1970s to the 1980s and again from the 1980s to the 1990s. We describe this analysis in the Supporting Information.

In addition to the historical analysis, we also examined the social factors that could underlie the degree of gender association in CS and CDS. In particular, we considered the socioeconomic status and racial background of a family. The Hall corpus, one of the largest subcorpora in CHILDES collected around the late 1970s–1980s, contains conversations tagged explicitly by the socioeconomic class and race of the families (see Supporting Information for details). It contained WC and MC families, including Black and White families in both classes. We performed the same analysis for gender probability in child–caretaker speech and gender associations in word embeddings, by splitting the Hall corpus data by socioeconomic class and race, respectively. Fig. 8 shows the strengths of the correlations between gender probability and word embedding associations by social class and race, using WEAT (see Supporting Information for a similar analysis using PROJ and other details). We find some

significant correlations between gender probability in speech and word embedding associations in both Black and White WC families, as correlations for 6 of 12 combinations of speech type, embeddings, and association test are significant with $p < .05$ for Black WC families and 7 of 12 for White WC families. Fewer cases are significant in MC families, as only 1 of 12 combinations shows a significant correlation ($p < .05$) in Black MC families and none shows significance for White MC families.

To compare across races, we pooled across socioeconomic classes (i.e., compared both MC and WC White families to MC and WC Black families), and likewise, we pooled across races to compare classes. Hypothesis tests of the difference between the correlation strengths across race and class groups mostly do not yield significance. Five of 12 combinations show a significant difference ($p < .05$) between WC and MC families, while 1 of 12 combinations show a difference between White and Black families. p-Values for the comparisons across race and class groups are Bonferroni-corrected to account for multiple comparisons. Complete results of these hypothesis tests are reported in the Supporting Information.

4. Discussion and conclusion

We have presented a large-scale quantitative investigation into the emergence of gender associations in child language development. Our emphasis is in quantifying gender associations through time as reflected in linguistic communication between young children and their caretakers. We have demonstrated the utility of word embeddings as a proxy for broad gender associations that helps to characterize gendered patterns in CS through both developmental and historical time courses. Our method for measuring gender associations incorporates information that is not captured sufficiently in existing methods for analyzing language development. Supporting Information provides evidence suggesting that these gender effects in language development are complementary to the psycholinguistic variables of word length, frequency, concreteness, and valence, which are common word metrics in developmental research (e.g., Braginsky, Yurovsky, Marchman, & Frank, 2019).

Our approach advances the study of gender associations in language. Previous studies have analyzed the strength of gender associations in speech by training word embeddings on corpora of interest and applying association tests to those embeddings. This approach requires a very large set of corpora to obtain reliable embeddings, such as the entire CHILDES corpus (e.g., Charlesworth et al., 2021). Our method uses a combination of pre-trained word embeddings and a simple metric of a word's relative probability in speech to draw inferences from smaller corpora. This alternative approach supports fine-grained analyses that stratify the CHILDES corpus by age, decade, and socioeconomic status. This methodology also helps to assess gender associations in CS under an external measure of those associations. As such, we can go beyond only pointing out gender differences in who says which words to whom by linking those differences to external measures that capture gender associations in the broader community.

Our investigation is based on the analyses of English-speaking children and caretakers. Although corpora in other languages are available in CHILDES, the scales of those corpora

are substantially smaller than in English. In addition, languages with grammatical gender pose a challenge for the analysis of gender associations, because the gender association of a word might reflect both its semantic association and grammatical gender. Extending our investigation toward a longer time span will also be constructive. Since the second wave of the American feminist movement began in the 1960s, it is possible that there was a large change in gender associations between the 1950s, 1960s, and 1970s. However, the CHILDES dataset is very sparse for the historical period prior to 1970.

One intriguing aspect of our findings is that the correlations tend to be weaker for fastText embeddings used with PROJ than for other pairs of word embeddings and association tests. In general, our results might vary between word embeddings due to differences in the corpora they were trained on, different optimization algorithms being used, or differences in specific techniques used, like the use of global word statistics in GloVe. Furthermore, this particular finding could be due to the interaction of two features of PROJ and fastText. While WEAT normalizes the strength of a word's gender association by dividing the difference in mean cosine similarities by the standard deviation of all similarities, PROJ does not. fastText represents words using a mixture of word-level representations and subwords. Ethayarajh et al. (2019) have shown that the normalization in WEAT can lead to overestimating the extent of gender associations when all words are close to each other, but likewise if many words share subword representations, they will be closer in fastText embedding space than in Word2Vec and GloVe embedding space. This suggests that the lack of normalization may lead gender associations to appear weaker in fastText compared to other embeddings.

Our study leaves open the question of whether the strength of the correlation between direct gender associations in speech and word embedding associations differs reliably between children's linguistic environment and that of adults. Our analysis of the Switchboard corpus revealed correlations of a similar strength to those in children between the ages of 3 and 5, but our analysis of the Santa Barbara corpus revealed weaker correlations. It is likely that the strength of these correlations differs by the situation in which conversations occur as well as by the age of their participants. The Santa Barbara Corpus contains adult conversations in naturalistic settings, while the Switchboard corpus contains telephone conversations. Many of the transcripts in the CHILDES corpus consist of children playing with toys, so gender differences in which toys children tend to play with could manifest in the CHILDES conversations. More generally, the extent to which the correlations observed between gender probability and word embedding associations generalize to all contexts in which children and caretakers speak with each other is still unclear. Thorough studies focused on how gendered information appears in different conversational settings would be valuable to understand whether the differences we see between the Santa Barbara Corpus and the CHILDES and Switchboard corpora are due to differences in conversational settings and how applicable our findings are to different contexts in which children and caretakers interact.

We extend the rich literature on language development and gender by suggesting that gender associations present in society have a clear imprint on child development, reflected in both linguistic input (caretakers' speech) and output (children's speech) in childhood, but the degree of gender association in CS may be modulated depending on social factors and changes. Our methodology connects machine learning to basic research in child

sociolinguistic development, and it creates future opportunities for probing the relations between natural language use and social biases through time.

Acknowledgments

We thank Claire Bowern for constructive comments, Aotao Xu for suggestions about the data, and Michael Frank for pointers. This work was supported by a Research Grant from the Institute for Gender and the Economy. BP is funded by an NSERC USRA award. SWSL is funded by an Early Researcher Award (Ontario Ministry of Research, Innovation and Science), an Insight Grant (SSHRC), and an Insight Development Grant (SSHRC). SS is funded through NSERC grant RGPIN-2017-06506. YX is funded through NSERC Discovery Grant RGPIN-2018-05872, SSHRC Insight Grant #435190272, and Ontario Early Researcher Award #ER19-15-050.

Notes

1 We recognize that "communication" is a complex activity that can encompass more than the text of the corpora we are analyzing, such as facial expression and gesture. For simplicity of expression, here we use the term "linguistic communication" to refer to the words used in communication to and by children, as reflected in our text corpora.
2 The exception is the combination of CS, fastText, and PROJ, where the p-value derived from bootstrapping and Bonferroni correction is exactly .05
3 To estimate how well t-SNE preserves the word similarity information in the word embeddings, we measured the correlation between the similarities of the full-dimensional and t-SNE-reduced word embeddings. We found that this method captured a significant proportion of the variance in the similarities of the high-dimensional word embeddings, though still a minority. We describe these results in Supporting Information.

Open Research Badges

 This article has earned Open Materials badge. Materials are available at https://osf.io/635em/?view_only=2870124df1a0467bbd6d032cf154551e.

References

Aubrey, J. S., & Harrison, K. (2004). The gender-role content of children's favorite television programs and its links to their gender-related perceptions. *Media Psychology*, *6*, 111–146.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160.

Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, *18*(3), 299–303.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In G. Tesauro, D. S. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems* (pp. 4349–4357). Cambridge, MA: MIT Press.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

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

Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). *Convokit: A toolkit for the analysis of conversations*. arXiv preprint arXiv:2005.04246.

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.

Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.

Donnelly, K., Twenge, J. M., Clark, M. A., Shaikh, S. K., Beiler-May, A., & Carter, N. T. (2016). Attitudes toward women's work and family roles in the united states, 1976–2013. *Psychology of Women Quarterly*, *40*(1), 41–54.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000). *Santa Barbara Corpus of Spoken American English*. CD-ROM. Philadelphia: Linguistic Data Consortium.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2019). Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American Psychologist*. *75*(3), 301–315.

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*, 275–298.

Endendijk, J. J., Groeneveld, M. G., Van der Pol, L. D., Van Berkel, S. R., Hallers-Haalboom, E. T., Mesman, J., & Bakermans-Kranenburg, M. J. (2014). Boys don't play with dolls: Mothers' and fathers' gender talk during picture book reading. *Parenting*, *14*, 141–161.

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Fagot, B. I., Leinbach, M. D., & O'boyle, C. (1992). Gender labeling, gender stereotyping, and parenting behaviors. *Developmental Psychology*, *28*, 225–230.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. Cambridge, MA: MIT Press.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 1 (pp. 517–520). IEEE Computer Society.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning word vectors for 157 languages*. arXiv preprint arXiv:1802.06893.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Hall, K., & Bucholtz, M. (2012). *Gender articulated: Language and the socially constructed self*. New York: Routledge.

Hall, W., Nagy, W., Linn, R., & Bruce, B. (1984). *Spoken words, effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Lawrence Associates.

Hall, W. S., Nagy, W. E., & Nottenburg, G. (1981). *Situational variation in the use of internal state words*. Center for the Study of Reading Technical Report no. 212.

Hall, W. S., & Tirre, W. C. (1979). *The communicative environment of young children: Social class, ethnic, and situational differences*. Center for the Study of Reading Technical Report no. 125.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change*. arXiv preprint arXiv:1605.09096.

Ehrlich, S., Meyerhoff, M., & Holmes, J. (2014). *The handbook of language, gender, and sexuality*. Hoboken, NJ: John Wiley & Sons.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236–248.

Kantrowitz, M. (1991). Name corpus: List of male, female, and pet names. https://www.cs.cmu.edu/Groups/AI/areas/nlp/corpora/names/0.html.

Lakoff, R. (1973). Language and woman's place. *Language in Society*, *2*(1), 45–79.

Laserna, C. M., Seih, Y.-T., & Pennebaker, J. W. (2014). Um… who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*(3), 328–338.

Leaper, C., Anderson, K. J., & Sanders, P. (1998). Moderators of gender effects on parents' talk to their children: A meta-analysis. *Developmental Psychology*, *34*(1), 3–27.

Lenton, A. P., Sedikides, C., & Bruder, M. (2009). A latent semantic analysis of gender stereotype-consistency and narrowness in American English. *Sex Roles*, *60*(3–4), 269–278.

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

Lovas, G. S. (2011). Gender and patterns of language development in mother-toddler and father-toddler dyads. *First Language*, *31*(1), 83–108.

Luther, C. A., & Legg Jr, J. R. (2010). Gender differences in depictions of social and physical aggression in children's television cartoons in the US. *Journal of Children and Media*, *4*, 191–205.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. New York: Psychology Press.

Martin, C., & Dinella, L. (2001). Gender-related development. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 6020–6027). Oxford: Pergamon.

Master, A., Meltzoff, A. N., & Cheryan, S. (2021). Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *Proceedings of the National Academy of Sciences*, *118*(48).

Meyer, M., & Gelman, S. A. (2016). Gender essentialism in children and parents: Implications for the development of gender stereotyping and gender-typed preferences. *Sex Roles*, *75*(9–10), 409–421.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In G. Tesauro, D. S. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems* (pp. 3111–3119). Cambridge, MA: MIT Press.

Mulac, A., Bradac, J. J., & Mann, S. K. (1985). Male/female language differences and attributional consequences in children's television. *Human Communication Research*, *11*, 481–506.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211–236.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Raag, T. (1999). Influences of social expectations of gender, gender stereotypes, and situational constraints on children's toy choices. *Sex Roles*, *41*, 809–831.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., Ungar, L. H., & Preis, T. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, *8*(9), e73791.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*, 672–686.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*, 659–673.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*, 654–666.

von der Malsburg, T., Poppels, T., & Levy, R. P. (2020). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 United States and 2017 United Kingdom elections. *Psychological Science*, *31*(2), 115–128.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1: Words with the highest gender probability in CDS. The higher the gender probability, the more the word is said disproportionately to girls

Table S2: Words with the lowest gender probability in CDS. The lower the gender probability, the more the word is said disproportionately to boys

Table S3: Words with gender probability closest to 0.5, which reflects the word being said equally to boys and girls

Table S4: Mean ρ across bootstrapped sub-samples of the CHILDES corpus for each combination of speech type, word embeddings, and association type. $p < .01$ for all correlations

Figure S1: Aggregate correlations between associations in word embeddings and speech using odds ratio

Figure S2: Aggregate correlations between associations in word embeddings and speech using log-odds ratio

Figure S3: Correlation strengths in Santa Barbara Corpus using gender probability to quantify gender associations in speech

Figure S4: Correlation strengths in Santa Barbara Corpus using odds ratio to quantify gender associations in speech

Figure S5: Correlation strengths in Santa Barbara Corpus using log-odds ratio to quantify gender associations in speech

Figure S6: Correlation strengths in the Switchboard Corpus using gender probability to quantify gender associations in speech

Figure S7: Correlation strengths in the Switchboard Corpus using odds ratio to quantify gender associations in speech

Figure S8: Correlation strengths in the Switchboard Corpus using log-odds ratio to quantify gender associations in speech

Table S5: Summary of hypothesis testing results for each year of child development across speech types, word embeddings, and association tests

Figure S9: Correlations between full-dimensional and t-SNE-reduced word embeddings

Figure S10: Visualization of words in child speech that show high and low gender probabilities, for age groups 1 and 5 in development

Figure S11: Correlation strengths of gender probability in child-directed and child speech with word embedding gender association, across socioeconomic status (working class, or WC, versus middle class, or MC) and race (black vs. white)

Table S6: Summary of hypothesis test results for each combination of race and social class across speech types, word embeddings, and tests. All p-values are Bonferroni-corrected to account for multiple comparisons

Table S7: Summary of pooled hypothesis test results between races and social classes across speech types, word embeddings, and tests

Table S8: Summary of hypothesis test results between pairs of decades across speech types, word embeddings, and tests

Figure S12: Correlations between gender probability in child-directed speech (left) and child speech (right) and gender associations in word embeddings based on PROJ (upper row) and WEAT (bottom row) tests

Table S9: Correlations between gender probability and psycholinguistic variables in child-directed speech (CDS) and child speech (CS)

Table S10: Coefficients from linear regression using psycholinguistic correlates of gender probability in CDS and CS

Table S11: Full and partial correlations between word embedding associations and gender probability. $p < .001$ in all cases. Partial correlations control for length, log-frequency, concreteness, and valence