The emergence of gender associations in child language development (Supporting Information)

Ben Prystawski[1,*], Erin Grant[2], Aida Nematzadeh[3], Spike W. S. Lee[4,5], Suzanne Stevenson[6], and Yang Xu[6,7]

[1]Department of Psychology, Stanford University, Stanford, California, USA
[2]Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, California, USA
[3]DeepMind, London, United Kingdom
[4]Department of Psychology, University of Toronto, Toronto, Ontario, Canada
[5]Rotman School of Management, University of Toronto, Toronto, Ontario, Canada
[6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[7]Cognitive Science Program, University of Toronto, Toronto, Ontario, Canada
[*]Correspondence: benpry@stanford.edu

Contents

List of Figures

List of Tables

1 List of explicitly gendered words

In addition to names and proper nouns, we removed all explicitly gendered words, such as "aunt" or "male," from our analyses. We also removed common nouns that could be names (like "frank" and "violet"). This list was composed by manually inspecting each word that occurs above the frequency threshold in any of the bootstrapped iterations of the corpus and adding that word and its corresponding equivalent (like "aunt" and "uncle"). The full list is as follows:

"aunt", "baba", "ballerina", "ballerino", "blond", "blonde", "boy", "boyfriend", "boys", "brother", "brothers", "bull", "cowboy", "cowboys", "cowgirl", "cowgirls", "dad", "dada", "dadda", "daddie", "daddies", "daddy", "dadee", "dads", "daisy", "daughter", "derrick", "don", "duchess", "dude", "duke", "emperor", "empress", "father", "female", "firemen", "frank", "gal", "gay", "gentlemen", "girl", "girlfriend", "girls", "godfather", "godmother", "gramma", "grampa", "grandfather", "grandma", "grandmother", "grandpa", "guy", "he", "hen", "her", "hers", "herself", "him", "himself", "his", "husband", "jack", "jackinthebox", "jill", "kiki", "king", "kings", "lad", "ladies", "lady", "lord", "ma", "madam", "mailman", "mailwoman", "male", "mam", "mama", "mami", "man", "men", "mia", "miss", "missus", "mister", "mom", "momee", "momma", "mommies", "mommy", "moms", "mother", "mum", "mummie", "nana", "pa", "papi", "penis", "peter", "policeman", "policemen", "policewoman", "policewomen", "poppop", "prince", "princess", "queen", "queens", "she", "sir", "sister", "sisters", "snowman", "son", "stepbrothers", "stepfather", "stepmother", "stepsisters", "steward", "stewardess", "superman", "uncle", "violet", "wife", "witch", "witches", "woman", "women".

2 Words with highest, lowest, and moderate gender probability in child-directed speech

To convey a sense of which words have high and low gender probabililty, we present lists of the words with the highest gender probability (Table S1), the lowest gender probability (Table S2), and gender probability closest to the midpoint of 0.5 (Table S3) in child-directed speech. All words in these tables appeared in the CDS portion of the CHILDES corpus at least 20 times. Some intuitive qualitative trends emerge from observing these words. Words said most disproportionately to girls often involve traditional girls' fashion ("tutu" and "pigtails") and toys ("dolly"). Many of the words said disproportionately to boys relate to violence and action ("violent," "bomber", and "pirate"). Words close to the midpoint include many function words, like "yes," "no," "on," and "all."

word	gender probability
creamcheese	1.00
giddy	1.00
nom	1.00
ponies	1.00
dale	1.00
tapioca	1.00
oompapa	1.00
pigtails	1.00
tinkertoy	1.00
tutu	1.00
puttaputta	1.00
dice	1.00
sleeper	1.00
yall	0.99
courage	0.97
mane	0.96
dolly	0.94
marry	0.94
marmalade	0.94
valentine	0.93
mash	0.93
stool	0.92
cottage	0.92
mam	0.92
untill	0.91
kittys	0.91
ponytail	0.91
cricket	0.91
fishie	0.90
ribbon	0.90

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

word	gender probability
ee	0.00
violent	0.00
moomilk	0.00
pirate	0.00
choochoo	0.00
shishi	0.00
clop	0.00
pau	0.00
spear	0.00
cemetery	0.00
budleyley	0.00
chugga	0.00
swingie	0.00
ooaa	0.00
scales	0.00
budleyleys	0.00
underoos	0.00
ninight	0.00
twerp	0.00
badada	0.00
didldow	0.00
squirrelie	0.00
bomber	0.00
nuuw	0.00
mwuh	0.00
shore	0.00
badji	0.00
infinity	0.00
nem	0.00
eeat	0.00

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

word	gender probability
unless	0.50
will	0.50
on	0.50
show	0.50
potato	0.50
writing	0.50
able	0.50
hamburger	0.50
me	0.50
eaten	0.50
please	0.50
mailman	0.50
first	0.50
there	0.50
nothing	0.50
in	0.50
pages	0.50
no	0.50
yes	0.50
anything	0.50
phone	0.50
tastes	0.50
than	0.50
mud	0.50
know	0.50
froggie	0.50
everything	0.50
all	0.50
awhile	0.50
yours	0.50

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

3 Means and confidence intervals of aggregate correlations

Table S4 shows the mean correlation strength between gender probability and word embedding associations across all 10,000 bootstrapped sub-samples of the corpus, as well as 95% confidence intervals. All correlations are highly significant, as $p < .01$ in all cases even after Bonferroni correction for multiple comparisons across different speech types (CS, CDS), word embedding types (Word2Vec, GloVe, fastText), and association tests (WEAT, PROJ).

Type	Embeddings	Test	Mean ρ [CI]
CDS	Word2Vec	WEAT	.18 [.16, .20]
CDS	Word2Vec	PROJ	.20 [.18, .22]
CDS	GloVe	WEAT	.24 [.22, .26]
CDS	GloVe	PROJ	.24 [.22, .26]
CDS	fastText	WEAT	.23 [.21, .26]
CDS	fastText	PROJ	.09 [.08, .11]
CS	Word2Vec	WEAT	.22 [.20, .24]
CS	Word2Vec	PROJ	.22 [.20, .24]
CS	GloVe	WEAT	.28 [.27, .30]
CS	GloVe	PROJ	.28 [.26, .30]
CS	fastText	WEAT	.25 [.24, .27]
CS	fastText	PROJ	.09 [.07, .10]

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.

4 Analyses with odds and log-odds ratios as alternative metrics

We describe additional analyses for evaluating the robustness of our findings, by using odds ratios and log-odds ratios as alternative metrics for the strength of gender associations in speech instead of gender probability as described in the main text. Odds ratio is defined as follows:

$$\frac{c(w, f)/c(\neg w, f)}{c(w, m)/c(\neg w, m)} \tag{1}$$

Here, $c(w, f)$ is the number of times word w is said to a female child and $c(\neg w, f)$ is the number of times all words other than w are said to a female child. Likewise, $c(w, m)$ is the number of times word w is said to a male child and $c(\neg w, m)$ is the number of times words other than w were said to a male child.

The log-odds ratio is the natural logarithm of the odds ratio as defined in Equation 1. The log-odds ratio results were very similar to the results based on gender probability as reported in the main text. Across all bootstrapped sub-samples and speech types, the average Pearson correlation between the gender probability and log-odds ratio of words is .79 ($p < .001$), which suggests that our results are robust to the choice of metric. The strength of the correlation between the odds ratio and gender probability is weaker although significant, averaging .29 ($p < .001$) across sub-samples and speech types. Both odds ratio and log-odds ratio yielded significant correlations with word embedding associations.

Figure S1 shows the aggregate correlations between odds ratios and word embedding associations in both child speech and child-directed speech, while Figure S2 shows the aggregate correlations between log-odds ratios and word embedding associations.

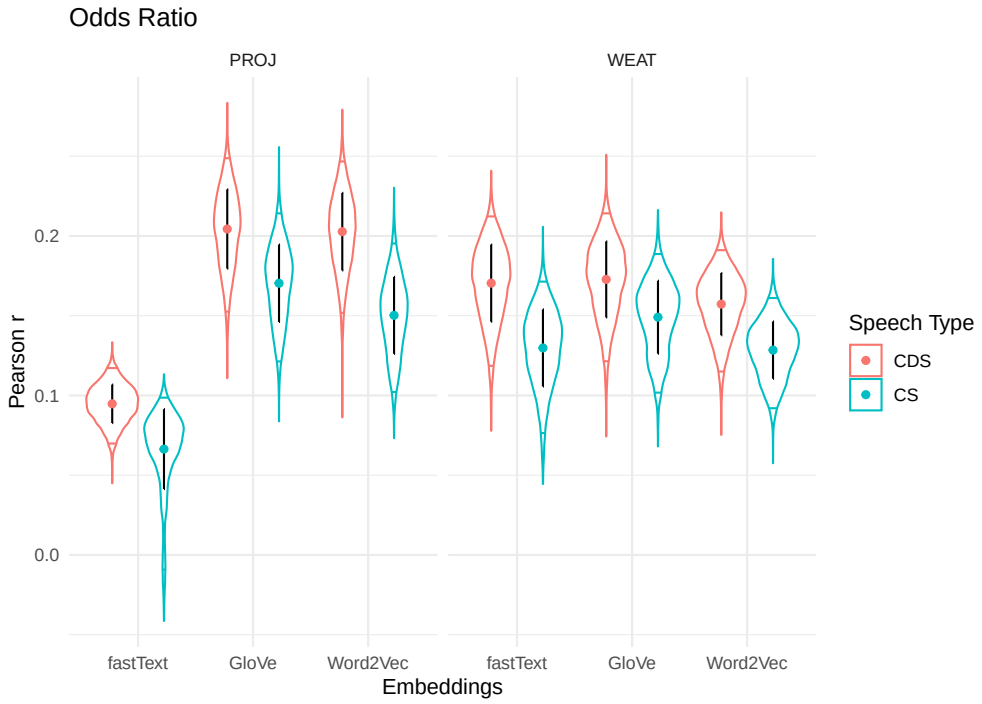

Figure S1: Aggregate correlations between associations in word embeddings and speech using odds ratio. WEAT stands for Word Embedding Association Test and PROJ refers to the Subspace Projection method, both of which are described in the main text. Raindrop plots show the density of correlation strengths 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.

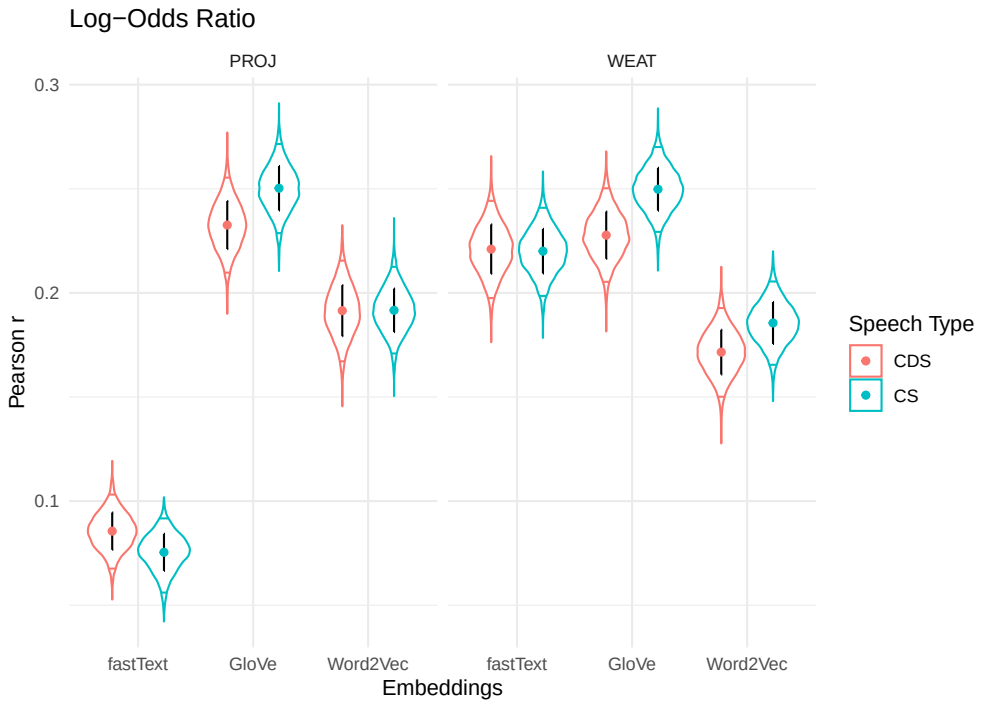

Figure S2: Aggregate correlations between associations in word embeddings and speech using log-odds ratio. WEAT stands for Word Embedding Association Test and PROJ refers to the Subspace Projection method, both of which are described in the main text. Raindrop plots show the density of correlation strengths 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.

5 Analyses of adult-adult speech

We analyzed adult-adult speech using two corpora: the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000) and the Switchboard corpus (Godfrey et al., 1992).

The Santa Barbara Corpus contains 340,860 tokens of speech between adults. Since this corpus is not tagged by the gender of speakers, we looked up the names of speakers in lists of predominantly male and predominantly female names (Kantrowitz, 1991) and used those tags as the gender labels. After removing speech by speakers whose names did not appear in the lists, we had 151,996 tokens of speech by men and 121,497 tokens of speech by women, for a total of 273,493 tokens. The correlations between gender probability in adult speech and word embedding associations are shown in Figure S3. The same analyses using odds ratio and log-odds ratio are shown in Figure S4 and Figure S5, respectively.

The Switchboard corpus contains 1,531,972 tokens of adult-adult telephone conversations. Speakers in this corpus are explicitly tagged by sex. The correlations are shown in Figure S6 (gender probability), Figure S7 (odds ratio), and Figure S8 (log-odds ratio).

We applied the same bootstrapping technique for our corpora of adult speech as we used for CHILDES. The bootstrapped sub-samples of the Santa Barbara corpus had a mean of 943 words above the frequency threshold of 20. The Switchboard corpus is larger, so it had a mean of 2,879 words above the threshold. On average, 668 of these words were shared between both corpora of adult speech, child speech in CHILDES, and child-directed speech in CHILDES. An average of 1,195 words were shared between the Switchboard corpus and CDS and CS in CHILDES. The number of non-overlapping words between corpora suggests that differences between the results in CHILDES and corpora of adult-adult speech could be due to differences in which words are used in speech involving children and speech between adults.

In general, correlations were stronger in the Switchboard corpus than in the Santa Barbara corpus. This could be because the Switchboard corpus is much larger and thus the word-level metrics are less noisy. It might also be due to inaccuracies in the gender tagging we did for the Santa Barbara corpus.

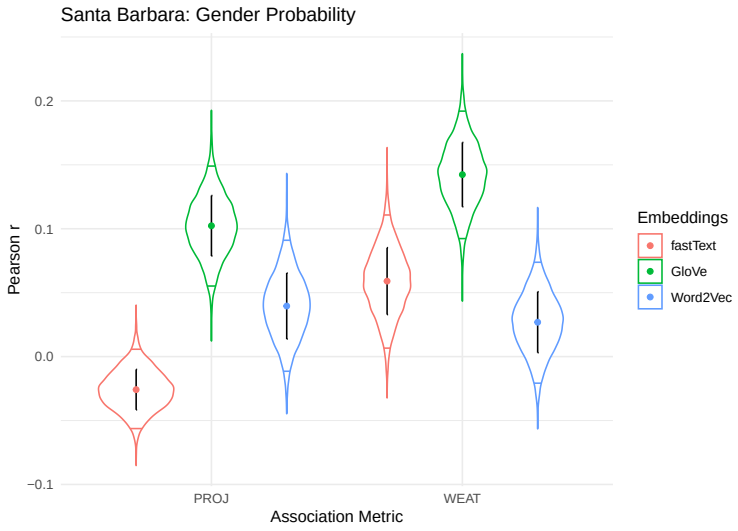

Figure S3: Correlation strengths in the Santa Barbara Corpus using gender probability to quantify gender associations in speech. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the Santa Barbara Corpus. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

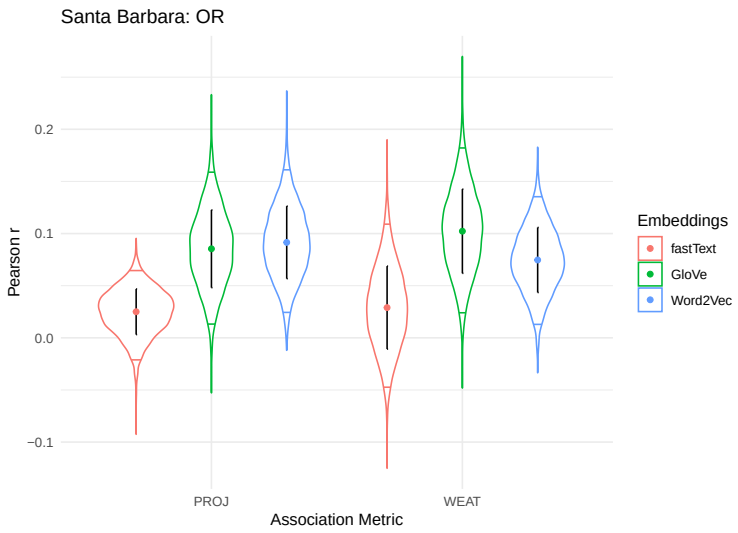

Figure S4: Correlation strengths in the Santa Barbara Corpus using odds ratio to quantify gender associations in speech. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the Santa Barbara Corpus. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

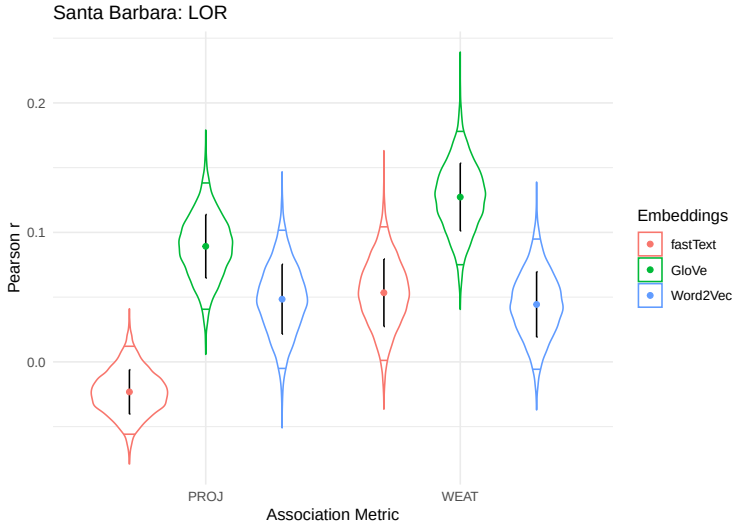

Figure S5: Correlation strengths in the Santa Barbara Corpus using log-odds ratio to quantify gender associations in speech. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the Santa Barbara Corpus. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

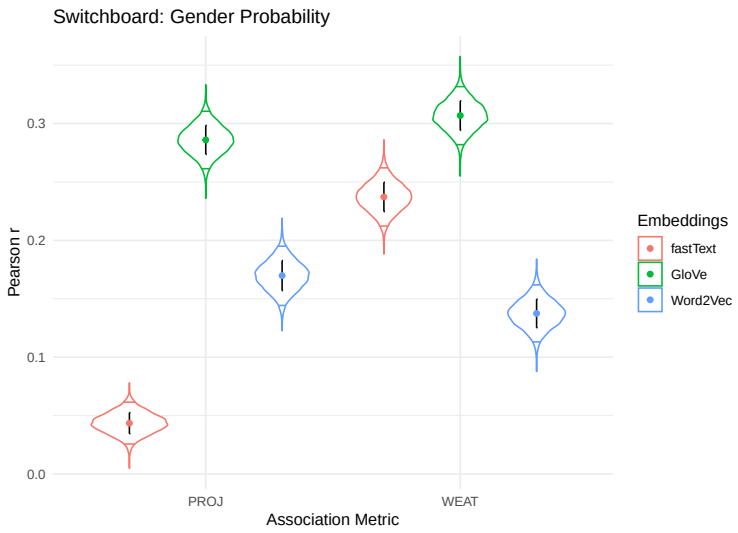

Figure S6: Correlation strengths in the Switchboard Corpus using gender probability to quantify gender associations in speech. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the Switchboard corpus. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

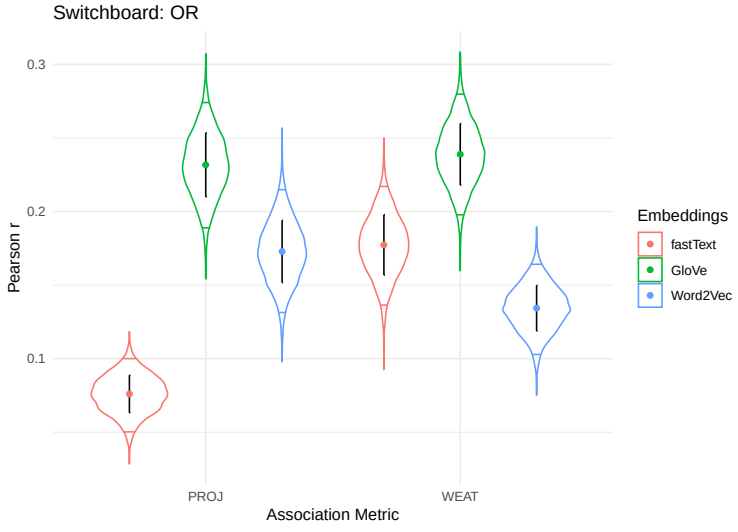

Figure S7: Correlation strengths in the Switchboard Corpus using odds ratio to quantify gender associations in speech. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the Switchboard corpus. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

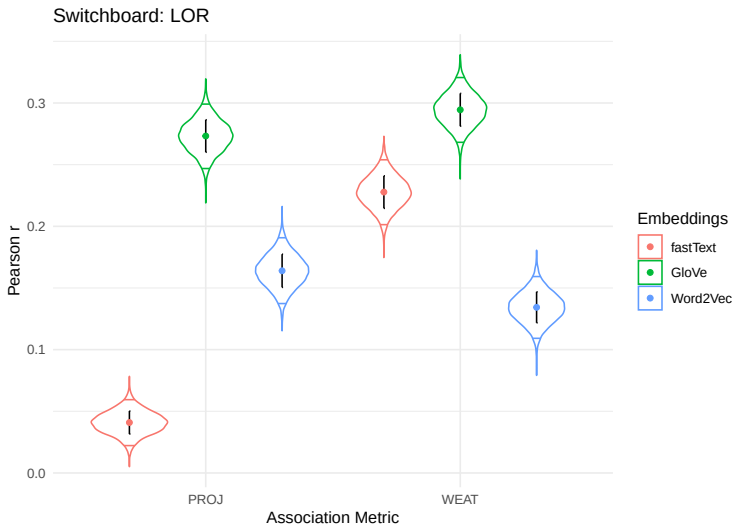

Figure S8: Correlation strengths in the Switchboard Corpus using log-odds ratio to quantify gender associations in speech. Raindrop plots show the density of correlation strengths across the 10,000 bootstrapped subsamples of the Switchboard corpus. Point estimates show the mean correlation across subsamples. Error bars denote standard error of the mean.

6 Hypothesis tests for analysis by child age

Table S5 shows the mean correlation strengths, p-values, and 95% confidence intervals for the analysis by child age. These statistics were computed from 10,000 bootstrapped subsamples of the CHILDES corpus, balanced by age and gender. Results are shown for each combination of speech type, word embeddings, and association test. p-values are Bonferroni-corrected to account for each of these combinations.

Table S5: Summary of hypothesis testing results for each year of child development across speech types, word embeddings, and association tests. All p-values are Bonferroni-corrected to account for multiple comparisons.

Type	Embeddings	Test	Age 1 ρ (p) [CI]	Age 2 ρ (p) [CI]	Age 3 ρ (p) [CI]	Age 4 ρ (p) [CI]	Age 5 ρ (p) [CI]
CDS	Word2Vec	WEAT	.00 (1.0) [-.04, .06]	.14 (< .01) [.10, .20]	.14 (< .01) [.10, .18]	.07 (.048) [.03, .11]	.17 (< .01) [.11, .22]
CDS	Word2Vec	PROJ	.02 (1.0) [-.02, .07]	.19 (< .01) [.14, .27]	.20 (< .01) [.15, .27]	.07 (.32) [.03, .13]	.17 (< .01) [.09, .22]
CDS	GloVe	WEAT	.06 (.62) [.01, .10]	.18 (< .01) [.13, .22]	.19 (< .01) [.13, .24]	.07 (1.0) [.01, .12]	.17 (< .01) [.10, .22]
CDS	GloVe	PROJ	.06 (.43) [.02, .10]	.18 (< .01) [.13, .22]	.22 (< .01) [.17, .26]	.09 (.14) [.03, .14]	.16 (< .01) [.09, .20]
CDS	fastText	WEAT	.05 (1.0) [.00, .09]	.14 (< .01) [.09, .19]	.20 (< .01) [.11, .26]	.09 (.77) [.01, .14]	.15 (< .01) [.06, .21]
CDS	fastText	PROJ	.02 (1.0) [-.02, .07]	.07 (< .01) [.04, .11]	.07 (< .01) [.05, .10]	.01 (1.0) [-.02, .04]	.08 (.02) [.05, .11]
CS	Word2Vec	WEAT	-.05 (1.0) [-.11, .03]	.17 (< .01) [.09, .22]	.21 (< .01) [.13, .27]	.17 (< .01) [.09, .22]	.22 (< .01) [.14, .28]
CS	Word2Vec	PROJ	-.04 (1.0) [-.09, .04]	.21 (< .01) [.14, .25]	.25 (< .01) [.16, .29]	.17 (< .01) [.09, .22]	.22 (< .01) [.14, .27]
CS	GloVe	WEAT	-.01 (1.0) [-.07, .07]	.23 (< .01) [.10, .30]	.27 (< .01) [.15, .34]	.22 (< .01) [.12, .28]	.26 (< .01) [.15, .33]
CS	GloVe	PROJ	-.04 (1.0) [-.09, .05]	.22 (< .01) [.10, .28]	.30 (< .01) [.19, .36]	.22 (< .01) [.12, .28]	.25 (< .01) [.15, .31]
CS	fastText	WEAT	-.02 (1.0) [-.08, .08]	.20 (< .01) [.09, .26]	.25 (< .01) [.13, .32]	.17 (< .01) [.07, .22]	.23 (< .01) [.13, .29]
CS	fastText	PROJ	-.04 (1.0) [-.09, .04]	.08 (.27) [.02, .11]	.07 (.02) [.05, .11]	.05 (.38) [.02, .09]	.11 (.050) [.06, .14]

7 Correlation between full-dimensional and t-SNE-reduced word embedding similarities

To examine how much of the variance in word embedding similarities is captured by the t-SNE-reduced vectors shown in Figure 6 of the main text, we measured the correlation between the pairwise similarities of the t-SNE-reduced vectors and the full-dimensional vectors over 10,000 random runs of t-SNE.

On each iteration, we initialized and fit two-dimensional t-SNE, then projected the word embeddings for each of the 60 words used in the plot in Figure 6 down to two dimensions. We computed pairwise cosine similarities between each pair of words in the set for both the full vectors and the reduced vectors, then computed the Pearson correlation between the cosine similarities. This tells us how accurately the similarities between the reduced vectors capture the similarities between the full-dimensional vectors.

As Figure S9 shows, there are significant positive correlations between full and reduced word embeddings. The average Pearson correlation is .19 across word embedding types. Our method captures a small amount of the variance in the full-dimensional vectors, but the portion captured is still statistically significant ($p < .01$ for all embeddings) and meaningful considering that we are projecting the embeddings down from around 300 dimensions to 2 dimensions. Still, our use of t-SNE-reduced vectors is primarily intended as an intuitive supplement to the quantitative analyses presented in other sections of the paper.

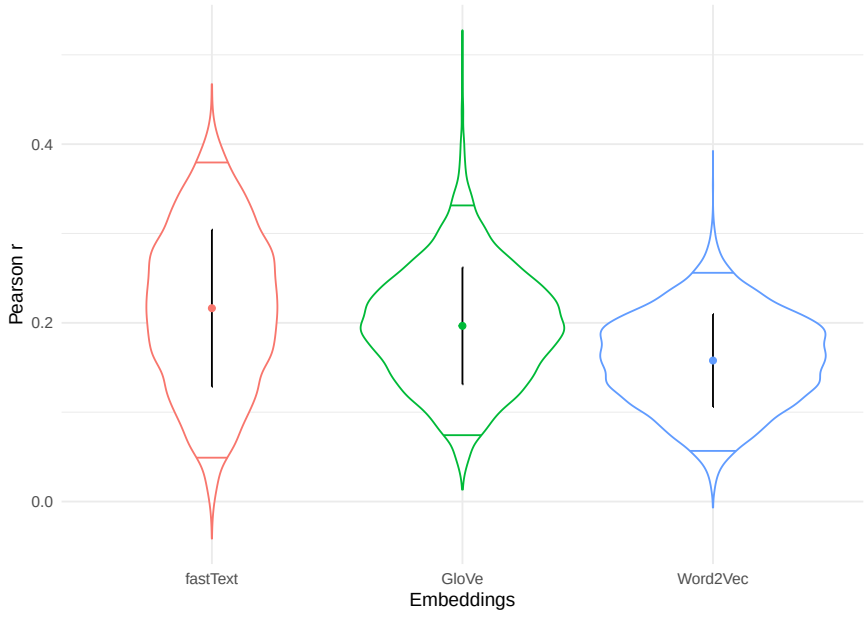

Figure S9: Correlations between full-dimensional and t-SNE-reduced word embeddings. Raindrop plots show the density of correlation strengths over 10,000 random runs of t-SNE. Point estimates show the mean correlation and error bars denote standard error of mean.

8 Dimension-reduced visualization of gender association in child speech

Figure S10 contains plots of t-SNE-reduced word embeddings, colored by their gender probability in child speech at ages 1 and 5. Words with high and low gender probability are largely similar between CS and CDS.

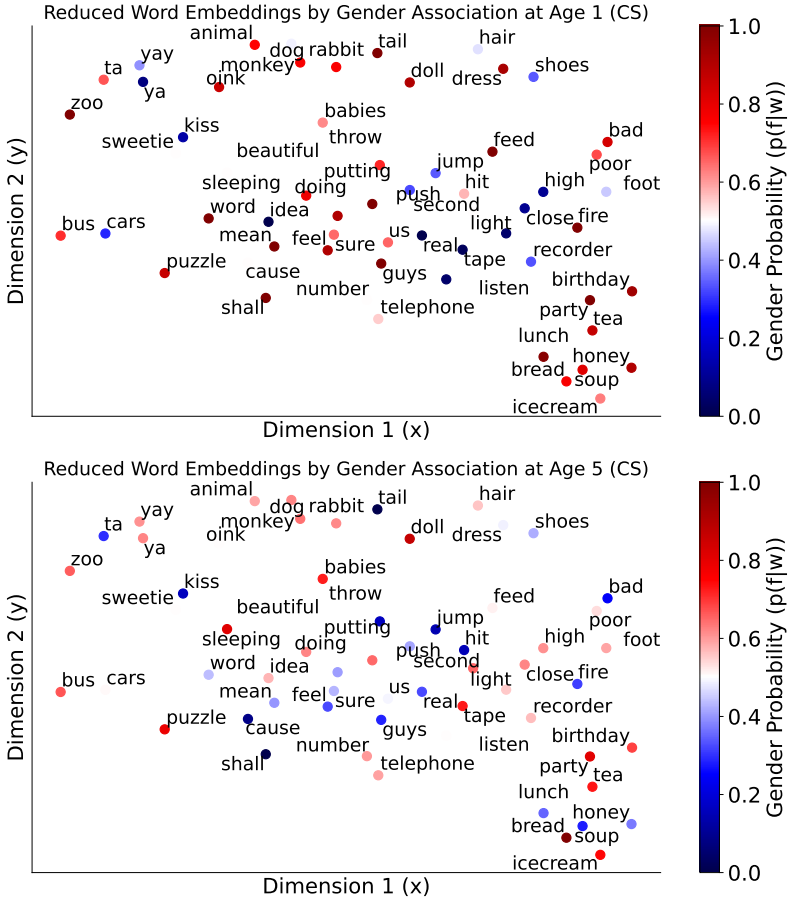

Figure S10: Visualization of words in child speech that show high and low gender probabilities, for age groups 1 and 5 in development. Semantic space is constructed from dimensionality-reduced Word2Vec embeddings. Colorbar indicates the scale of gender probability, with 1 indicating words exclusively uttered by girls and 0 exclusively by boys.

9 Analyses of social class and race with subspace projection

Figure S11 summarizes the results on social factors of gender association using the Subspace Projection method (PROJ). Corresponding hypothesis tests are reported in Section 10.

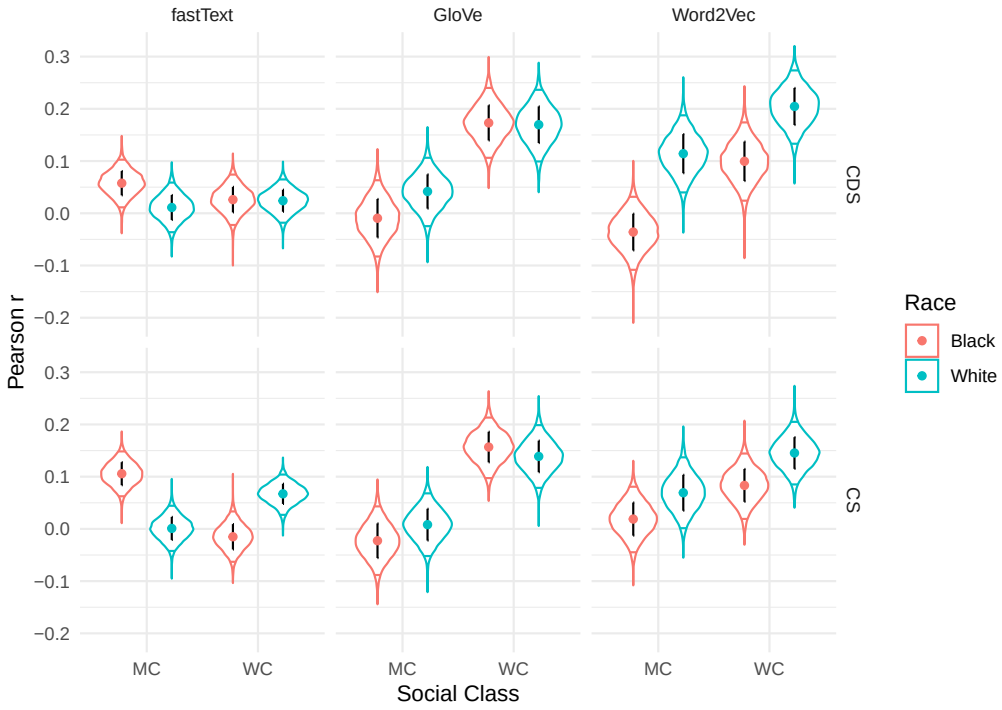

Figure S11: Correlation strengths of gender probability in child-directed and child speech with word-embedding gender association, across socio-economic status (working class, or WC, versus middle class, or MC) and race (Black vs. White). Raindrop plots show the density of correlation strengths across 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 PROJ.

10 Hypothesis tests for social class and race

Here, we report the results of hypothesis tests comparing the strength of the correlations across social class and race groups. Table S6 shows the hypothesis test results for each pair of race and socioeconomic class. p-values for these tests are Bonferroni-corrected to account for multiple comparisons across each combination of race, socio-economic class, speech type, association test, and word embeddings. Table S7 shows the results of hypothesis tests of the differences between classes and socioeconomic statuses. For these tests, correlation strengths were averaged over the other variable (e.g. we average over classes when comparing races). These p-values are Bonferroni-corrected to account for multiple comparisons. Many of these results are statistically insignificant, so it is inconclusive whether there are significant differences in the extent of gender associations in child speech by the race and socio-economic class of a family.

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.

Type	Embeddings	Test	Black, WC d (p) [CI]	Black, MC d (p) [CI]	White, WC d (p) [CI]	White, MC d (p) [CI]
CDS	Word2Vec	WEAT	.11 (.29) [.03, .19]	-.02 (1.0) [-.08, .05]	.15 (< .01) [.08, .22]	.05 (1.0) [-.02, .13]
CDS	Word2Vec	PROJ	.10 (.41) [.02, .17]	-.04 (1.0) [-.11, .03]	.20 (< .01) [.13, .27]	.11 (.15) [.04, .19]
CDS	GloVe	WEAT	.20 (< .01) [.11, .27]	.00 (1.0) [-.08, .08]	.13 (.05) [.05, .21]	.05 (1.0) [-.03, .12]
CDS	GloVe	PROJ	.17 (< .01) [.11, .24]	-.01 (1.0) [-.08, .06]	.17 (< .01) [.10, .24]	.04 (1.0) [-.02, .11]
CDS	fastText	WEAT	.19 (< .01) [.11, .27]	-.03 (1.0) [-.11, .04]	.08 (1.0) [-.0, .16]	.06 (1.0) [-.02, .13]
CDS	fastText	PROJ	.03 (1.0) [-.02, .07]	.06 (.63) [.01, .10]	.02 (1.0) [-.02, .06]	.01 (1.0) [-.04, .06]
CS	Word2Vec	WEAT	.09 (.27) [.03, .16]	.01 (1.0) [-.05, .07]	.11 (< .01) [.05, .18]	.06 (1.0) [.00, .13]
CS	Word2Vec	PROJ	.08 (.40) [.02, .14]	.02 (1.0) [-.04, .08]	.15 (< .01) [.09, .20]	.07 (1.0) [.00, .14]
CS	GloVe	WEAT	.13 (< .01) [.07, .20]	.01 (1.0) [-.06, .08]	.13 (< .01) [.06, .20]	.04 (1.0) [-.03, .10]
CS	GloVe	PROJ	.16 (< .01) [.10, .21]	-.02 (1.0) [-.09, .04]	.14 (< .01) [.08, .20]	.01 (1.0) [-.05, .07]
CS	fastText	WEAT	.12 (< .01) [.06, .19]	.06 (1.0) [-.01, .12]	.09 (.45) [.02, .15]	.01 (1.0) [-.06, .07]
CS	fastText	PROJ	-.02 (1.0) [-.06, .03]	.11 (< .01) [.06, .15]	.07 (.07) [.03, .10]	.00 (1.0) [-.04, .04]

Table S7: Summary of pooled hypothesis test results between races and social classes across speech types, word embeddings, and tests. All p-values are Bonferroni-corrected to account for multiple comparisons.

Type	Embeddings	Test	WC vs. MC d (p) [CI]	White vs. Black d (p) [CI]
CDS	Word2Vec	WEAT	.11 (.06) [.04, .18]	-.06 (1.0) [-.13, .02]
CDS	Word2Vec	PROJ	.11 (.046) [.04, .18]	-.13 (.01) [-.2, -.06]
CDS	GloVe	WEAT	.14 (< .01) [.06, .22]	.01 (1.0) [-.07, .08]
CDS	GloVe	PROJ	.16 (< .01) [.09, .22]	-.02 (1.0) [-.09, .04]
CDS	fastText	WEAT	.12 (.06) [.05, .20]	.01 (1.0) [-.07, .09]
CDS	fastText	PROJ	-.01 (1.0) [-.06, .04]	.02 (1.0) [-.02, .07]
CS	Word2Vec	WEAT	.06 (1.0) [.00, .13]	-.04 (1.0) [-.1, .03]
CS	Word2Vec	PROJ	.07 (.72) [.01, .13]	-.06 (1.0) [-.12, .01]
CS	GloVe	WEAT	.11 (.046) [.04, .18]	-.01 (1.0) [-.08, .05]
CS	GloVe	PROJ	.16 (< .01) [.10, .22]	-.01 (1.0) [-.07, .05]
CS	fastText	WEAT	.07 (.77) [.01, .14]	.04 (1.0) [-.02, .11]
CS	fastText	PROJ	-.03 (1.0) [-.07, .02]	.01 (1.0) [-.03, .05]

11 Additional information on evaluation of changes by decade

We performed hypothesis tests comparing the mean strength of gender associations across bootstrapped iterations between the 70s and 80s, 80s and 90s, and 70s and 90s. We did this for each combination of speech type (CDS or CS), word embeddings (Word2Vec, GloVe, or fastText), and test type (WEAT or PROJ). Table S8 summarizes the effect sizes and p-values for each decade. p-values are Bonferroni-corrected to account for multiple comparisons.

Table S8: Summary of hypothesis test results between pairs of decades across speech types, word embeddings, and tests. Positive effect sizes correspond to a decrease from one decade to the next while negative effect sizes correspond to an increase. All p-values are Bonferroni-corrected to account for multiple comparisons.

Type	Embeddings	Test	d (p) [CI] 70s-80s	d (p) [CI] 80s-90s	d (p) [CI] 70s-90s
CDS	Word2Vec	WEAT	.09 (< .01) [.06, .13]	.12 (< .01) [.08, .16]	.21 (< .01) [.17, .25]
CDS	Word2Vec	PROJ	.11 (< .01) [.08, .15]	.12 (< .01) [.08, .16]	.23 (< .01) [.19, .27]
CDS	GloVe	WEAT	.14 (< .01) [.10, .18]	.06 (.18) [.02, .11]	.20 (< .01) [.16, .24]
CDS	GloVe	PROJ	.16 (< .01) [.12, .19]	.07 (.04) [.03, .11]	.23 (< .01) [.19, .26]
CDS	fastText	WEAT	.12 (< .01) [.08, .16]	-.02 (1.0) [-.06, .03]	.10 (< .01) [.06, .14]
CDS	fastText	PROJ	.11 (< .01) [.08, .14]	-.02 (1.0) [-.05, .01]	.09 (< .01) [.06, .12]
CS	Word2Vec	WEAT	.03 (1.0) [-.01, .07]	.18 (< .01) [.13, .22]	.20 (< .01) [.17, .24]
CS	Word2Vec	PROJ	.06 (.14) [.02, .09]	.14 (< .01) [.10, .18]	.20 (< .01) [.16, .24]
CS	GloVe	WEAT	.02 (1.0) [-.01, .06]	.14 (< .01) [.10, .18]	.17 (< .01) [.13, .21]
CS	GloVe	PROJ	.06 (.047) [.02, .10]	.11 (< .01) [.07, .15]	.17 (< .01) [.13, .21]
CS	fastText	WEAT	.02 (1.0) [-.02, .06]	.10 (< .01) [.06, .14]	.12 (< .01) [.08, .16]
CS	fastText	PROJ	.03 (1.0) [-.00, .05]	.02 (1.0) [-.01, .05]	.05 (.0504) [.02, .08]

12 Analyses by decade with diachronic word embeddings

In addition to the three sets of word embeddings used in the main analysis, we also analyzed correlation strengths by decade using historical word embeddings. Since the ways words are used in general language changes over time, it could be possible that differences between decades are the result of gender norms changing relative to modern word embeddings rather than a trend toward weaker gender associations. To rule out this possibility, we used HistWords embeddings, which are trained independently on text from each decade between the 1800s and 1990s (Hamilton et al., 2016). These embeddings are available online here: https://nlp.stanford.edu/projects/histwords/.

For each decade we studied, we compared gender probability in speech against word embedding associations in the HistWords embeddings for the corresponding decade. This means that, for example, speech from the 1970s is compared against word embeddings trained on text from the 1970s. Figure S12 shows the strength of correlations between gender probability in CDS and CS and word embedding associations using the HistWords embeddings.

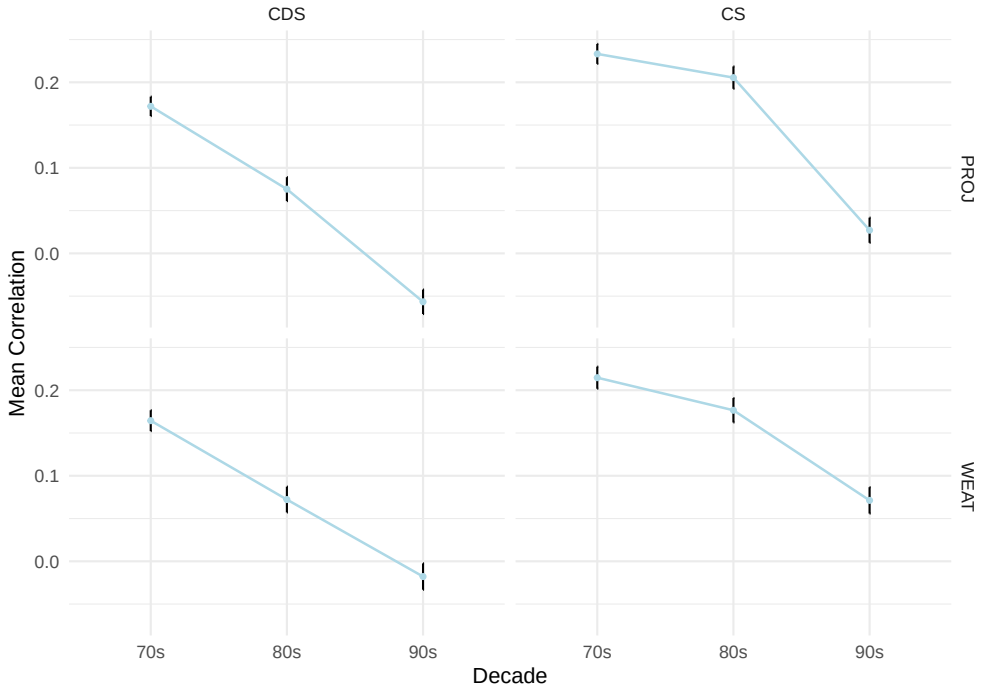

Figure S12: Correlations between gender probability in child-directed speech (left) and child speech (right) and gender associations in diachronic word embeddings from the corresponding decade based on PROJ (upper row) and WEAT (bottom row) tests. Point estimates denote mean correlation strengths across 10,000 bootstrapped subsamples of the corpus and error bars denote standard error of the mean.

13 Psycholinguistic correlates of gender probability

We examined the extent to which our measure of gender probability might be explained by established psycholinguistic variables in the literature of child language development. For this analysis, we used all corpora except for the Hall corpus and did not use bootstrapping. We considered four common word metrics from the developmental literature (e.g., Braginsky et al., 2019): word form length, usage frequency, concreteness, and valence. We computed the length of each word and estimated the frequency in child-directed speech and child speech directly from the CHILDES data by counting the occurrences of the word said to or by all children respectively. We took concreteness and valence ratings of words from existing large-scale behavioral experiments in Warriner et al. (2013) and Brysbaert et al. (2014). In these experiments, humans rated word concreteness or valence from 1-10. Ratings were averaged across participants. We then measured the correlation between gender probability and each of the four variables. Data for word concreteness are available at http://crr.ugent.be/archives/1330. Data for word valence are available at http://crr.ugent.be/archives/1003.

Table S9 summarizes the Pearson correlation coefficients from this analysis. We analyzed only words which occurred at least 20 times in the corpus and found significant but small ($\rho \leq 0.14$) correlations between the gender probability of a word and the psycholinguistic metrics that we considered. In particular, we found that words with shorter length, higher frequency, more positive valence, and higher concreteness tend also to be said more to and by girls than boys. The correlation between gender probability in child-directed speech and valence is consistent with the previous finding from Leaper et al. (1998) that mothers use more supportive language when speaking to girls compared with boys. We applied linear regression to all four psycholinguistic variables to account for gender probability in both child-directed speech and child speech. The coefficients with 95% confidence intervals are summarized in Table S10. These variables together explain the variance in gender probability to a good degree, with R^2 values being 0.630 for child-directed speech and 0.582 for child speech. We computed the partial correlation between gender probabilities and word embedding gender associations while controlling for the four psycholinguistic variables. We focused on the 955 words for which this data is available in the datasets from Warriner et al. (2013) and Brysbaert et al. (2014). Results of this analysis are summarized in Table S11. Controlling for these factors only reduces the correlation strength by .04-.05 ($p < .001$ for the partial correlations in all cases). These results show that the gender probabilities in child development are both correlative and complementary to the other factors, because the variability in gender effects cannot be explained solely by the psycholinguistic variables that we considered here.

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

Variable	Pearson ρ	p-value
Length (CDS)	.073	.0012
Length (CS)	$-.033$	.23
Log-frequency (CDS)	.069	.0025
Log-frequency (CS)	.13	$< .001$
Concreteness (CDS)	.05	.022
Concreteness (CS)	.083	.0023
Valence (CDS)	.14	$< .001$
Valence (CS)	.14	$< .001$

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

Variable	CS	CDS
Length	-.0007 [-.006, .005]	.0168 [.013, .020]
Log frequency	.0172 [.011, .023]	.0161 [.012, .021]
Concreteness	.0488 [.039, .058]	.0325 [.024, .038]
Valence	.0315 [.024, .039]	.0286 [.028, .040]

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.

Embedding	CS	CS (partial)	CDS	CDS (partial)
Word2Vec	.261	.238	.243	.211
GloVe	.327	.311	.290	.260
fastText	.285	.262	.260	.225

References

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

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

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

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

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

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.

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

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.